Model-Free Variable Selection through Sufficient Dimension Reduction

---

A Dissertation
Submitted to
the Temple University Graduate Board

---

in Partial Fulfillment
of the Requirements for the Degree of
DOCTOR OF PHILOSOPHY

---

by
Angela Minster
May, 2016

ABSTRACT

Model-Free Variable Selection through Sufficient Dimension Reduction

Angela Minster

DOCTOR OF PHILOSOPHY

Temple University, May, 2016

Professor Yuexiao Dong, Chair

In this thesis we draw upon the natural connection between the fields of sufficient dimension reduction and variable selection to develop new theory and methods for model-free variable selection. After developing the natural connection between sufficient dimension reduction and model-free variable selection we introduce two approaches to select independent variables important to predicting the response variable without making any assumptions about the function form of the relationship between predictor and response. The first is a stepwise procedure and the second takes a penalized approach. Both are rooted in ordinary least squares regression but with modifications to facilitate model-free variable selection. We also introduce a set of transformations for model-free variable selection. Finally we develop a stepwise procedure that is able to select interaction terms in the model-free setting. We show the effec-

tiveness of these methods through simulation studies and an analysis of real data.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

1.1 Overview

When creating a statistical model to explain a phenomenon, correctly selecting the key explanatory variables is of the utmost importance. Thus, research on variable selection has been a common theme in the recent literature. Classical variable selection methods include stepwise linear regression and best subset selection. See, for example, Thompson (1978). As data grow larger and the number of predictor variables from which to select increases, many researchers have necessarily sought to improve upon classical methods. One key advance has been the introduction of penalized regression methods. Since the introduction of LASSO (Tibshirani, 1996), many penalized least squares methods have been developed, such as SCAD (Fan and Li, 2001), elastic net

(Zou and Hastie, 2005), adaptive LASSO (Zou, 2006), Dantzig selector (Candes and Tao, 2007), MCP (Zhang, 2010), etc. These aforementioned methods are all model-based, and they all rely on the linear model assumption. Since the linear model assumption can be easily violated in application, more flexible variable selection procedures are needed to fill the gap.

Our proposed methods are based on the concept of Sufficient Dimension Reduction (SDR) (Li, 1991; Cook, 1994). For response $Y \in \mathbb{R}$ and predictor $\mathbf{x} = (X_1, \ldots, X_p)^T \in \mathbb{R}^p$, let $F(Y \mid \mathbf{x})$ denote the conditional distribution function of Y given $\mathbf{x}$. One particular goal of SDR is to find indices $\boldsymbol{\beta}_1^T\mathbf{x}, \cdots, \boldsymbol{\beta}_q^T\mathbf{x}$, where $\boldsymbol{\beta}_i = (\beta_{i1}, \ldots, \beta_{ip})^T \in \mathbb{R}^p$, $i = 1, \ldots, q$ and $q < p$, such that

$$F(Y \mid \mathbf{x}) = F(Y \mid \boldsymbol{\beta}_1^T\mathbf{x}, \cdots, \boldsymbol{\beta}_q^T\mathbf{x}). \tag{1.1}$$

Thus the distribution of Y depends on $\mathbf{x}$ only through the q indices $\boldsymbol{\beta}_1^T\mathbf{x}, \cdots, \boldsymbol{\beta}_q^T\mathbf{x}$. On the other hand, the goal of model-free variable selection is to find an active predictor index set $\mathcal{A}_0$ such that

$$F(Y \mid \mathbf{x}) = F(Y \mid \mathbf{x}_{\mathcal{A}_0}). \tag{1.2}$$

Here $\mathbf{x}_{\mathcal{A}_0} = \{X_k \colon k \in \mathcal{A}_0\}$ contains all the predictors whose subscripts are contained in the set $\mathcal{A}_0$, and (1.2) implies that the distribution of Y depends on $\mathbf{x}$ only through the predictors contained in $\mathbf{x}_{\mathcal{A}_0}$. Let $\Omega = \{1, 2, \cdots, p\}$ be the full index set. It follows from (1.2) that we can define active set $\mathcal{A}_0$ as

$$\mathcal{A}_0 = \{k \in \Omega \colon F(Y \mid \mathbf{x}) \text{ depends on } X_k\}. \tag{1.3}$$

From the similar forms of (1.1) and (1.2), one can immediately see the following important connection between SDR and model-free variable selection. On one hand, if $k \in \mathcal{A}_0$, or X_k is active in terms of predicting $F(Y \mid \mathbf{x})$, then $\beta_{ik} \neq 0$ for at least one of $i = 1, \ldots, q$. On the other hand, if $k \notin \mathcal{A}_0$, or X_k is inactive in terms of predicting $F(Y \mid \mathbf{x})$, then $\beta_{ik} = 0$ for all of $i = 1, \ldots, q$.

Through this important connection, it becomes obvious that one can recover the active set $\mathcal{A}_0$ directly through the estimation of of $\boldsymbol{\beta}_i$, $i = 1, \ldots, q$. However, direct estimation of $\boldsymbol{\beta}_i$ for the purpose of model-free variable selection may be problematic for the following reasons. First, there are numerous popular SDR methods in the literature, including as sliced inverse regression (SIR) (Li, 1991), sliced average variance estimation (SAVE) (Cook and Weisberg, 1991), directional regression (Li and Wang, 2007), ordinary least squares (OLS) (Duan and Li, 1989), principal Hessian directions (PHD) (Li, 1992; Cook, 1998), etc. It is not clear which method should be recommended for the purpose of model-free variable selection. Second, certain distribution assumptions are required for each of the aforementioned SDR methods, and variable transformations are common practice in SDR to facilitate such distribution assumptions. See, for example, predictor transformation in Cook and Nachtsheim (1994) and response transformation in Yin and Cook (2002). So while transformation is well studied in the SDR literature, it is not immediately clear what effect variable transformation will have on model-free variable

selection. Last but not least, in the age of Big Data, direct estimation of $\boldsymbol{\beta}_i$ for all $i = 1, \ldots, q$ in the $p > n$ setting is at best computationally challenging and time-consuming and at worst not even possible. It is of great computational advantage if one can achieve model-free variable selection without estimating all q indices $\boldsymbol{\beta}_1^T\mathbf{x}, \cdots, \boldsymbol{\beta}_q^T\mathbf{x}$.

This dissertation aims to address all of the issues raised above, and proposes effective tools for model-free variable selection through SDR. Our main contributions are four-fold. First, we propose the principle of using OLS for model-free variable selection. Unlike slicing-based SDR methods such as SIR and SAVE, which can lead to different results based on different choices of slice numbers, OLS does not require slicing of the response. It is well-known in the SDR literature that OLS is a limited method for the purpose of estimating indices $\boldsymbol{\beta}_i$'s as OLS is guaranteed to fail if there are more than one index ($q \geq 2$). So while OLS can estimate at most one index under the SDR framework we reveal that multi-index models are generally not a problem when using OLS for the purpose of model-free variable selection. As a matter of fact, estimating one index even if $q \geq 2$ can greatly simply the computation. For example, existing algorithms such as LASSO (Tibshirani, 1996) and Dantzig selector (Candes and Tao, 2007) can be readily used for OLS, and the computation for model-free variable selection thus becomes straightforward in the case of

$p > n$. The principle of using OLS for model-free variable selection is described in Section 2.

Second, the effectiveness of using OLS for model-free variable selection depends on certain assumptions. In the cases when such assumptions are violated, variable transformations are necessary before application of any OLS-based variable selection algorithms. Specifically, we discuss three types of variable transformations. In the case when the predictor distribution assumption is violated, marginal predictor normalization is proposed. In the case when the response depends on the predictor through the heteroskedastic error, response transformation is proposed. In the case when the unknown regression link function between the response and the active predictor is symmetric, a polynomial transformation of the predictor is considered. The principle of variable transformation to facilitate model-free variable selection is studied in Section 3.

Third, two alternative algorithms for model-free variable selection are proposed under the general framework of OLS, and both algorithms can be applied together with the variable transformation paradigm discussed earlier. Parallel to the classical stepwise regression in linear regression, the first algorithm is a stepwise regression procedure based on a novel permutation test. The second algorithm is a direct application of the $L1$ penalized regression methods, which include Dantzig selector and LASSO. Although Dantzig selector and

LASSO are originally proposed for linear regression, we demonstrate that $L1$ penalized regression is also applicable for the purpose of model-free variable selection. Both the stepwise regression algorithm and the $L1$ penalized regression algorithm are applicable in the challenging case of $p > n$. The two algorithms are developed in Section 4, and the numerical studies including variable transformations are presented in Section 5.

The fourth contribution is concerns model-free interaction detection. For a simple interaction-only model $Y = X_1X_2 + \epsilon$, OLS fails to recover active predictors X_1 and X_2, and none of variable transformation procedures described before can help with interaction detection. We notice that PHD (Li, 1992; Cook, 1998) can recover $\boldsymbol{\beta}_1 = (1, 0, \ldots, 0)^T$ and $\boldsymbol{\beta}_2 = (0, 1, 0, \ldots, 0)^T$ in this interaction-only model. This motivates us to propose a novel stepwise regression algorithm based on PHD, which is parallel to the OLS-based stepwise regression algorithm in Section 4. The PHD-based algorithm for model-free interaction detection is described in Section 6.

1.2 Review of OLS and PHD for Sufficient Dimension Reduction

We briefly review OLS (Duan and Li, 1989) and PHD (Li, 1992; Cook, 1998) in this section. There are two types of PHD (principle Hessian direc-

tions) estimation methods, the response-based PHD and the residual-based PHD, and we focus on the response-based PHD. Unless specified otherwise, we assume $E(\mathbf{x}) = \mathbf{0}$ and $E(Y) = 0$ throughout this section.

Let $\mathbf{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_q) \in \mathbb{R}^{p\times q}$ and denote $\perp\!\!\!\perp$ as statistical independence. Equation (1.1) then implies that

$$Y \perp\!\!\!\perp \mathbf{x} \mid \mathbf{B}^T\mathbf{x}. \tag{1.4}$$

Note that if $\mathbf{B} \in \mathbb{R}^{p\times q}$ satisfies (1.4), then $\mathbf{BA}$ also satisfies (1.4) for any full-rank matrix $\mathbf{A} \in \mathbb{R}^{q\times q}$. Thus $\mathbf{B}$ is only unique up to its column space. For any $\mathbf{B}$ satisfying (1.4), we refer to the column space of $\mathbf{B}$ as a dimension reduction subspace. The intersection of all dimension reduction subspaces is called the central space, and is denoted by $\mathcal{S}_{Y|\mathbf{x}}$ (Cook, 1994). Without ambiguity, we assume that $\mathbf{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_q)$ is a basis of $\mathcal{S}_{Y|\mathbf{x}}$.

Let $\boldsymbol{\mu} = \mathrm{E}(\mathbf{x})$ and $\boldsymbol{\Sigma} = \mathrm{var}(\mathbf{x})$ be the mean vector and the covariance matrix of predictor $\mathbf{x}$ respectively. Then $\mathbf{z} = \boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})$ is the standardized predictor. An important property of the central space is the invariance property (Cook, 1998). For $p \times p$ nonsingular matrix $\mathbf{A}$ and constant vector $\mathbf{b} \in \mathbb{R}^p$, consider linear transformation $\mathbf{w} = \mathbf{Ax} + \mathbf{b}$. The invariance property states that $\mathcal{S}_{Y|\mathbf{w}} = \mathbf{A}^{-1}\mathcal{S}_{Y|\mathbf{x}}$. A practical implication of this property is that one can first work with the standardized predictor $\mathbf{z}$ to get $\mathcal{S}_{Y|\mathbf{z}}$, and then transform it to the $\mathbf{x}$-scale using $\mathcal{S}_{Y|\mathbf{x}} = \boldsymbol{\Sigma}^{-1/2}\mathcal{S}_{Y|\mathbf{z}}$.

Denote the OLS estimator as $\boldsymbol{\beta}_{\mathrm{OLS}} = \boldsymbol{\Sigma}^{-1}\mathrm{E}(\mathbf{x}Y)$. The next result suggests that OLS can be used for recovering the central space.

Proposition 1.1. *Suppose* $\mathbf{x}$ *is normal with* $E(\mathbf{x}) = \mathbf{0}$. *Then* $\boldsymbol{\beta}_{OLS} \subseteq \mathcal{S}_{Y|\mathbf{x}}$.

Proof. Let $\mathrm{span}(\mathbf{B})$ denote the column space of $\mathbf{B}$. For $\mathrm{span}(\mathbf{B}) = \mathcal{S}_{Y|\mathbf{x}}$, the conditional independence (1.4) implies that $\mathrm{E}(Y \mid \mathbf{x}) = \mathrm{E}(Y \mid \mathbf{B}^T\mathbf{x})$. We have

$$\mathrm{E}(\mathbf{x}Y) = \mathrm{E}\{\mathbf{x}\mathrm{E}(Y \mid \mathbf{x})\} = \mathrm{E}\{\mathbf{x}\mathrm{E}(Y \mid \mathbf{B}^T\mathbf{x})\} = \mathrm{E}\{\mathrm{E}(\mathbf{x} \mid \mathbf{B}^T\mathbf{x})Y\}.$$

Denote $\mathbf{P}_{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}\mathbf{B}(\mathbf{B}^T\boldsymbol{\Sigma}\mathbf{B})^{-1}\mathbf{B}^T$. The normality of $\mathbf{x}$ guarantees that

$$\mathrm{E}(\mathbf{x} \mid \mathbf{B}^T\mathbf{x}) = \mathbf{P}_{\boldsymbol{\Sigma}}\mathbf{x}. \tag{1.5}$$

It follows that $\mathrm{E}(\mathbf{x}Y) = \mathrm{E}(\mathbf{P}_{\boldsymbol{\Sigma}}\mathbf{x}Y) = \mathbf{P}_{\boldsymbol{\Sigma}}\mathrm{E}(\mathbf{x}Y) = \boldsymbol{\Sigma}\mathbf{B}(\mathbf{B}^T\boldsymbol{\Sigma}\mathbf{B})^{-1}\mathbf{B}^T\mathrm{E}(\mathbf{x}Y)$. Thus we have

$\boldsymbol{\beta}_{\mathrm{OLS}} = \boldsymbol{\Sigma}^{-1}\mathrm{E}(\mathbf{x}Y) = \mathbf{B}(\mathbf{B}^T\boldsymbol{\Sigma}\mathbf{B})^{-1}\mathbf{B}^T\mathrm{E}(\mathbf{x}Y) \subseteq \mathrm{span}(\mathbf{B}) = \mathcal{S}_{Y|\mathbf{x}}$. □

We remark that condition (1.5) is known as the linear conditional mean (LCM) condition, and the normality assumption in Proposition 1.1 can be replaced with the weaker LCM condition.

Denote $\boldsymbol{\Sigma}_{\mathrm{Yxx}} = \mathrm{E}(Y\mathbf{x}\mathbf{x}^T)$. The next result suggests that the column space of $\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_{\mathrm{Yxx}}$ belongs to the central space.

Proposition 1.2. *Suppose* $E(Y) = 0$ *and* $\mathbf{x}$ *is normal with* $E(\mathbf{x}) = \mathbf{0}$. *Then* $span(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_{\mathrm{Yxx}}) \subseteq \mathcal{S}_{Y|\mathbf{x}}$.

PROOF. For $\text{span}(\mathbf{B}) = \mathcal{S}_{Y|\mathbf{x}}$, the conditional independence (1.4) implies that $\text{E}(Y \mid \mathbf{x}) = \text{E}(Y \mid \mathbf{B}^T\mathbf{x})$. We have

$$\text{E}(Y\mathbf{x}\mathbf{x}^T) = \text{E}\{\text{E}(Y \mid \mathbf{x})\mathbf{x}\mathbf{x}^T\} = \text{E}\{\text{E}(Y \mid \mathbf{B}^T\mathbf{x})\mathbf{x}\mathbf{x}^T\} = \text{E}\{Y\text{E}(\mathbf{x}\mathbf{x}^T \mid \mathbf{B}^T\mathbf{x})\} \quad (1.6)$$

Denote $\mathbf{Q}_{\mathbf{\Sigma}} = (\mathbf{I} - \mathbf{P}_{\mathbf{\Sigma}})\mathbf{\Sigma}$. The normality of $\mathbf{x}$ guarantees (1.5) as well as

$$\text{var}(\mathbf{x} \mid \mathbf{B}^T\mathbf{x}) = \mathbf{Q}_{\mathbf{\Sigma}}. \quad (1.7)$$

Thus $\text{E}(\mathbf{x}\mathbf{x}^T \mid \mathbf{B}^T\mathbf{x}) = \text{var}(\mathbf{x} \mid \mathbf{B}^T\mathbf{x}) + \text{E}(\mathbf{x} \mid \mathbf{B}^T\mathbf{x})\text{E}^T(\mathbf{x} \mid \mathbf{B}^T\mathbf{x}) = \mathbf{Q}_{\mathbf{\Sigma}} + \mathbf{P}_{\mathbf{\Sigma}}\mathbf{x}\mathbf{x}^T\mathbf{P}_{\mathbf{\Sigma}}^T$. Plugging into (1.6), together with $\mathbf{\Sigma}_{\text{Y}\mathbf{x}\mathbf{x}} = \text{E}(Y\mathbf{x}\mathbf{x}^T)$ and $\text{E}(Y) = 0$, we get

$$\mathbf{\Sigma}_{\text{Y}\mathbf{x}\mathbf{x}} = \text{E}(Y\mathbf{x}\mathbf{x}^T) = \mathbf{Q}_{\mathbf{\Sigma}}E(Y) + \mathbf{P}_{\mathbf{\Sigma}}\text{E}(Y\mathbf{x}\mathbf{x}^T)\mathbf{P}_{\mathbf{\Sigma}}^T = \mathbf{P}_{\mathbf{\Sigma}}\mathbf{\Sigma}_{\text{Y}\mathbf{x}\mathbf{x}}\mathbf{P}_{\mathbf{\Sigma}}^T.$$

Note that $\mathbf{P}_{\mathbf{\Sigma}} = \mathbf{\Sigma}\mathbf{B}(\mathbf{B}^T\mathbf{\Sigma}\mathbf{B})^{-1}\mathbf{B}^T$, we have $\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_{\text{Y}\mathbf{x}\mathbf{x}} = \mathbf{B}(\mathbf{B}^T\mathbf{\Sigma}\mathbf{B})^{-1}\mathbf{B}^T\mathbf{\Sigma}_{\text{Y}\mathbf{x}\mathbf{x}}\mathbf{P}_{\mathbf{\Sigma}}^T$. It follows that $\text{span}(\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_{\text{Y}\mathbf{x}\mathbf{x}}) \subseteq \text{span}(\mathbf{B}) = \mathcal{S}_{Y|\mathbf{x}}$. □

We remark that condition (1.7) is known as the constant conditional variance (CCV) condition. From the proof of Proposition 1.2, we can see that the normality assumption can be replaced with the LCM condition (1.5) and the CCV condition (1.7). Proposition 1.2 provides the theoretical justification for PHD, which uses the left singular vectors corresponding to the nonzero singular values of $\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_{\text{Y}\mathbf{x}\mathbf{x}}$ to recover $\mathcal{S}_{Y|\mathbf{x}}$.

The next result provides the connection between OLS and PHD.

Proposition 1.3. *Suppose* $E(Y) = 0$ *and* $\mathbf{x}$ *is normal with* $E(\mathbf{x}) = \mathbf{0}$. *Then*

$$E\left\{\frac{\partial E(Y \mid \mathbf{x})}{\partial \mathbf{x}}\right\} = \boldsymbol{\Sigma}^{-1}E(\mathbf{x}Y) \text{ and } E\left\{\frac{\partial^2 E(Y \mid \mathbf{x})}{\partial \mathbf{x}\partial \mathbf{x}^T}\right\} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_{\mathrm{Yxx}}\boldsymbol{\Sigma}^{-1}.$$

Proof. From first-order Stein's Lemma, we have $\mathrm{E}\left\{\frac{\partial g(\mathbf{x})}{\partial \mathbf{x}}\right\} = \boldsymbol{\Sigma}^{-1}\mathrm{E}\{\mathbf{x}g(\mathbf{x})\}$. From second-order Stein's Lemma, we have $\mathrm{E}\left\{\frac{\partial^2 g(\mathbf{x})}{\partial \mathbf{x}\partial \mathbf{x}^T}\right\} = \boldsymbol{\Sigma}^{-1}\mathrm{E}\{g(\mathbf{x})(\mathbf{x}\mathbf{x}^T - \boldsymbol{\Sigma})\}\boldsymbol{\Sigma}^{-1}$. Take $g(\mathbf{x}) = \mathrm{E}(Y \mid \mathbf{x})$. Then we get the desired result by noticing that $\mathrm{E}\{\mathrm{E}(Y \mid \mathbf{x})\mathbf{x}\} = \mathrm{E}(Y\mathbf{x})$, $\mathrm{E}\{\mathrm{E}(Y \mid \mathbf{x})\mathbf{x}\mathbf{x}^T\} = \mathrm{E}(Y\mathbf{x}\mathbf{x}^T) = \boldsymbol{\Sigma}_{\mathrm{Yxx}}$ and $\mathrm{E}\{\mathrm{E}(Y \mid \mathbf{x})\boldsymbol{\Sigma}\} = \mathrm{E}(Y)\boldsymbol{\Sigma} = \mathbf{0}$. □

For $\mathrm{E}(Y \mid \mathbf{x})$ with unknown function form, Proposition 1.3 states that OLS is essentially using the direction information contained in the average first order derivative of $\mathrm{E}(Y \mid \mathbf{x})$, while PHD coincides with the direction information contained in the average second order derivative of $\mathrm{E}(Y \mid \mathbf{x})$. Consider model $Y = (X_1 + 2X_2)^2 + \epsilon$. Then $\mathrm{E}(Y \mid \mathbf{x}) = (\boldsymbol{\theta}^T\mathbf{x})^2$ with $\boldsymbol{\theta} = (1, 2, 0, \ldots, 0)^T$. It follows that $\frac{\partial \mathrm{E}(Y|\mathbf{x})}{\partial \mathbf{x}} = 2\boldsymbol{\theta}^T\mathbf{x}\boldsymbol{\theta}$, and $\boldsymbol{\beta}_{\mathrm{OLS}} = \mathrm{E}(2\boldsymbol{\theta}^T\mathbf{x}\boldsymbol{\theta}) = \mathbf{0}$. On the other hand, consider model model $Y = X_1 + 2X_2 + \epsilon$. Then $\frac{\partial \mathrm{E}(Y|\mathbf{x})}{\partial \mathbf{x}} = \boldsymbol{\theta}$ and $\frac{\partial^2 \mathrm{E}(Y|\mathbf{x})}{\partial \mathbf{x}\partial \mathbf{x}^T} = \mathbf{0}$. Thus $\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_{\mathrm{Yxx}}\boldsymbol{\Sigma}^{-1} = \mathbf{0}$ and $\mathrm{span}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_{\mathrm{Yxx}}) = \mathbf{0}$. We reveal that for the purpose of recovering $\mathcal{S}_{Y|\mathbf{x}}$, OLS does not work well when $\mathrm{E}(Y \mid \mathbf{x})$ is symmetric, and PHD does not work well when $\mathrm{E}(Y \mid \mathbf{x})$ is linear in $\mathbf{x}$.

1.3 Central Mean Space and Central Kth Moment Space

Recall that in the definition of central space $\mathcal{S}_{Y|\mathbf{x}}$, we aim to find linear combinations of $\mathbf{x}$ such that $Y \perp\!\!\!\perp \mathbf{x} \mid \mathbf{B}^T\mathbf{x}$, and we have $F(Y \mid \mathbf{x}) = F(Y \mid \mathbf{B}^T\mathbf{x})$ as a result. When the interest is about the conditional mean $\mathrm{E}(Y \mid \mathbf{x})$ instead of the conditional distribution $F(Y \mid \mathbf{x})$, Cook and Li (2002) propose to consider

$$Y \perp\!\!\!\perp \mathrm{E}(Y \mid \mathbf{x}) \mid \mathbf{B}^T\mathbf{x}. \tag{1.8}$$

If $\mathbf{B}$ satisfies (1.8), it is easy to see that $\mathrm{E}(Y \mid \mathbf{x}) = \mathrm{E}(Y \mid \mathbf{B}^T\mathbf{x})$. If $\mathbf{B} \in \mathbb{R}^{p\times q}$ satisfies (1.8), we refer to the column space of $\mathbf{B}$ as a mean dimension reduction subspace. The intersection of all mean dimension reduction subspaces is the smallest mean dimension reduction subspace. We call it the central mean space, and denote it by $\mathcal{S}_{\mathrm{E}(Y|\mathbf{x})}$.

Similar to the idea of central mean space, denote $M^{(k)}(Y \mid \mathbf{x}) = \mathrm{E}\{(Y - \mathrm{E}(Y \mid \mathbf{x}))^k \mid \mathbf{x}\}$ for $k \geq 2$ and $M^{(1)}(Y \mid \mathbf{x}) = \mathrm{E}(Y \mid \mathbf{x})$. Yin and Cook (2002) consider

$$Y \perp\!\!\!\perp \{M^{(1)}(Y \mid \mathbf{x}), \ldots, M^{(K)}(Y \mid \mathbf{x})\} \mid \mathbf{B}^T\mathbf{x}. \tag{1.9}$$

If $\mathbf{B}$ satisfies (1.9), it is easy to see that $M^{(k)}(Y \mid \mathbf{x}) = M^{(k)}(Y \mid \mathbf{B}^T\mathbf{x})$, for $k = 1, \ldots, K$. If $\mathbf{B} \in \mathbb{R}^{p\times q}$ satisfies (1.9), we refer to the column space of $\mathbf{B}$ as a Kth moment dimension reduction subspace. The intersection of all Kth

moment dimension reduction subspaces is the smallest Kth moment dimension reduction subspace. We call it the central Kth moment space, and denote it by $\mathcal{S}_{Y|\mathbf{x}}^{(K)}$.

By the definitions above, central mean space is a special case of the central Kth moment space with $K = 1$. Because $Y \perp\!\!\!\perp \mathbf{x} \mid \mathbf{B}^T\mathbf{x}$ implies $Y \perp\!\!\!\perp \mathrm{E}(Y \mid \mathbf{x}) \mid \mathbf{B}^T\mathbf{x}$, a dimension reduction subspace is also a mean dimension reduction subspace. It follows that $\mathcal{S}_{\mathrm{E}(Y|\mathbf{x})} \subseteq \mathcal{S}_{Y|\mathbf{x}}$, since the former is the intersection of at least the same, if not more, subspaces. The central Kth moment space is contained in the central space for the same reasoning. We can also see that a dimension reduction subspace is always an ith moment dimension reduction space, which must be a jth moment dimension reduction space for any $j \leq i$. These relationships can be summarized as below,

$$\mathcal{S}_{\mathrm{E}(Y|\mathbf{x})} = \mathcal{S}_{Y|\mathbf{x}}^{(1)} \subseteq \cdots \subseteq \mathcal{S}_{Y|\mathbf{x}}^{(K)} \subseteq \cdots \subseteq \mathcal{S}_{Y|\mathbf{x}}.$$

If the conditional distribution of Y given $\mathbf{x}$ depends only on up to the Kth moments of $\mathbf{x}$, then $\mathcal{S}_{Y|\mathbf{x}} = \mathcal{S}_{Y|\mathbf{x}}^{(K)}$.

The next result provides estimators in the central Kth moment space.

Proposition 1.4. *Suppose* $\mathbf{x}$ *is normal with* $E(\mathbf{x}) = \mathbf{0}$. *Then for* $k = 1, \ldots, K$, *we have*

$$\boldsymbol{\Sigma}^{-1}E(\mathbf{x}Y^k) \subseteq \mathcal{S}_{Y|\mathbf{x}}^{(K)} \text{ and } span\left(\boldsymbol{\Sigma}^{-1}E[\{(Y^k - E(Y^k)\}\mathbf{x}\mathbf{x}^T]\right) \subseteq \mathcal{S}_{Y|\mathbf{x}}^{(K)}.$$

PROOF. For $\text{span}(\mathbf{B}) = \mathcal{S}^{(K)}_{Y|\mathbf{x}}$, the conditional independence (1.9) implies that $\text{E}(Y^k \mid \mathbf{x}) = \text{E}(Y^k \mid \mathbf{B}^T\mathbf{x})$ for all $k \leq K$. Thus

$$\text{E}(\mathbf{x}Y^k) = \text{E}\{\mathbf{x}\text{E}(Y^k \mid \mathbf{x})\} = \text{E}\{\mathbf{x}\text{E}(Y^k \mid \mathbf{B}^T\mathbf{x})\} = \text{E}\{\text{E}(\mathbf{x} \mid \mathbf{B}^T\mathbf{x})Y^k\} = \mathbf{P}_{\mathbf{\Sigma}}\text{E}(\mathbf{x}Y^k).$$

It follows that $\mathbf{\Sigma}^{-1}\text{E}(\mathbf{x}Y^k) = \mathbf{B}(\mathbf{B}^T\mathbf{\Sigma}\mathbf{B})^{-1}\mathbf{B}^T\text{E}(\mathbf{x}Y^k) \subseteq \text{span}(\mathbf{B}) = \mathcal{S}^{(K)}_{Y|\mathbf{x}}$.

Because $\text{E}(Y^k \mid \mathbf{x}) = \text{E}(Y^k \mid \mathbf{B}^T\mathbf{x})$, we also have $\text{E}[\{(Y^k - \text{E}(Y^k)\}\mathbf{x}\mathbf{x}^T] = \text{E}[\{\text{E}\{Y^k - \text{E}(Y^k) \mid \mathbf{x}\}\mathbf{x}\mathbf{x}^T] = \text{E}[\{\text{E}\{Y^k - \text{E}(Y^k) \mid \mathbf{B}^T\mathbf{x}\}\mathbf{x}\mathbf{x}^T] = \text{E}[\{Y^k - \text{E}(Y^k)\}\text{E}(\mathbf{x}\mathbf{x}^T \mid \mathbf{B}^T\mathbf{x})]$. Plug in $\text{E}(\mathbf{x}\mathbf{x}^T \mid \mathbf{B}^T\mathbf{x}) = \mathbf{Q}_{\mathbf{\Sigma}} + \mathbf{P}_{\mathbf{\Sigma}}\mathbf{x}\mathbf{x}^T\mathbf{P}^T_{\mathbf{\Sigma}}$ and note that $\text{E}[\{(Y^k - \text{E}(Y^k)\}\mathbf{Q}_{\mathbf{\Sigma}}] = \mathbf{0}$, we get

$$\text{E}[\{(Y^k - \text{E}(Y^k)\}\mathbf{x}\mathbf{x}^T] = \mathbf{P}_{\mathbf{\Sigma}}\text{E}[\{(Y^k - \text{E}(Y^k)\}\mathbf{x}\mathbf{x}^T]\mathbf{P}^T_{\mathbf{\Sigma}}.$$

Note that $\mathbf{P}_{\mathbf{\Sigma}} = \mathbf{\Sigma}\mathbf{B}(\mathbf{B}^T\mathbf{\Sigma}\mathbf{B})^{-1}\mathbf{B}^T$, we have $\mathbf{\Sigma}^{-1}\text{E}[\{(Y^k - \text{E}(Y^k)\}\mathbf{x}\mathbf{x}^T] = \mathbf{B}(\mathbf{B}^T\mathbf{\Sigma}\mathbf{B})^{-1}\mathbf{B}^T\text{E}[\{(Y^k - \text{E}(Y^k)\}\mathbf{x}\mathbf{x}^T]\mathbf{P}^T_{\mathbf{\Sigma}}$. It follows that $\text{span}\left(\mathbf{\Sigma}^{-1}\text{E}[\{(Y^k - \text{E}(Y^k)\}\mathbf{x}\mathbf{x}^T]\right) \subseteq \text{span}(\mathbf{B}) = \mathcal{S}^{(K)}_{Y|\mathbf{x}}$. □

In the special case of $K = 1$, Proposition 1.4 implies that both the OLS and the PHD estimators belong the central mean space $\mathcal{S}_{\text{E}(Y|\mathbf{x})}$. If the conditional distribution $F(Y \mid \mathbf{x})$ depends on $\mathbf{x}$ only through the regression mean $\text{E}(Y \mid \mathbf{x})$, then $\mathcal{S}_{\text{E}(Y|\mathbf{x})} = \mathcal{S}_{Y|\mathbf{x}}$, and OLS and PHD are targeting the central space. If the conditional distribution $F(Y \mid \mathbf{x})$ depends on $\mathbf{x}$ through higher moments such as $\text{E}\{(Y - \text{E}(Y \mid \mathbf{x}))^2 \mid \mathbf{x}\}$, then OLS and PHD are estimating a proper subspace of the central space.

CHAPTER 2

Model-Free Variable Selection through OLS: The Principle

Recall that $\mathcal{A}_0 = \{k \in \Omega\colon F(Y \mid \mathbf{x}) \text{ depends on } X_k\}$ denotes the active predictor set, where $\Omega = \{1, 2, \cdots, p\}$ is the full index set. To describe the principle of using OLS for model-free variable selection, we make the following assumptions throughout this section:

Condition (a): $\mathbf{x}$ is normal with $\mathrm{E}(\mathbf{x}) = \mathbf{0}$ and $\mathrm{var}(\mathbf{x}) = \mathbf{\Sigma}$;

Condition (b): $\mathrm{E}\{\partial \mathrm{E}(Y \mid \mathbf{x})/\partial X_k\} \neq 0$ for any $k \in \mathcal{A}_0$;

Condition (c): $\mathcal{S}_{\mathrm{E}(Y|\mathbf{x})} = \mathcal{S}_{Y|\mathbf{x}}$.

2.1 The Hypotheses Testing Approach

For some working index set $\mathcal{F}$, let $\mathbf{x}_{\mathcal{F}} = \{X_k \colon k \in \mathcal{F}\}$. For a candidate index $j \in \mathcal{F}^c$, where $\mathcal{F}^c$ denotes the complement of $\mathcal{F}$ in Ω, consider the following hypotheses

$$H_0 \colon Y \perp\!\!\!\perp X_j \mid \mathbf{x}_{\mathcal{F}} \text{ v.s. } H_a \colon Y \text{ is not independent of } X_j \text{ given } \mathbf{x}_{\mathcal{F}}. \tag{2.1}$$

We refer to (2.1) as the stepwise regression hypotheses, which is used to test whether Y is independent of X_j conditional on $\mathbf{x}_{\mathcal{F}}$. The next result describes a key quantity $\delta_{j|\mathcal{F}}$ that can be used for testing the stepwise regression hypotheses (2.1).

Proposition 2.1. *Suppose condition (a) holds. Let* $\Psi_{\mathcal{F},j}(\mathbf{x}) = E(Y \mid \mathbf{x}_{\mathcal{F}}, X_j)$ *and* $\delta_{j|\mathcal{F}} = E(YR_{j|\mathcal{F}})$*, where* $R_{j|\mathcal{F}} = X_j - E(X_j \mid \mathbf{x}_{\mathcal{F}})$*. Then*

i. $\delta_{j|\mathcal{F}} = \mathrm{var}(R_{j|\mathcal{F}})E\{\partial\Psi_{\mathcal{F},j}(\mathbf{x})/\partial X_j\}$*;*

ii. $\delta_{j|\mathcal{F}} = 0$ *under* H_0 *in (2.1).*

Proof. For part i, Stein's Lemma implies that

$$\mathrm{E}\left\{ \frac{\partial\Psi_{\mathcal{F},j}(\mathbf{x})}{\partial X_j} \;\middle|\; \mathbf{x}_{\mathcal{F}} \right\} = \mathrm{var}^{-1}(X_j \mid \mathbf{x}_{\mathcal{F}})\mathrm{E}\{\Psi_{\mathcal{F},j}(\mathbf{x})(X_j - \mathrm{E}(X_j \mid \mathbf{x}_{\mathcal{F}})) \mid \mathbf{x}_{\mathcal{F}}\}. \tag{2.2}$$

Take expectation on both sides of (2.2) and we get

$$\mathrm{E}\{\partial\Psi_{\mathcal{F},j}(\mathbf{x})/\partial X_j\} = \mathrm{var}^{-1}(X_j \mid \mathbf{x}_{\mathcal{F}})\mathrm{E}\{Y(X_j - \mathrm{E}(X_j \mid \mathbf{x}_{\mathcal{F}}))\}. \tag{2.3}$$

The normality of $\mathbf{x}$ implies that

$$\mathrm{E}(X_j \mid \mathbf{x}_{\mathcal{F}}) = \mathrm{E}^T(X_j\mathbf{x}_{\mathcal{F}})\boldsymbol{\Sigma}_{\mathcal{F}}^{-1}\mathbf{x}_{\mathcal{F}}, \text{ where } \boldsymbol{\Sigma}_{\mathcal{F}} = \mathrm{var}(\mathbf{x}_{\mathcal{F}}). \tag{2.4}$$

Then $R_{j|\mathcal{F}} = X_j - E^T(X_j\mathbf{x}_{\mathcal{F}})\boldsymbol{\Sigma}_{\mathcal{F}}^{-1}\mathbf{x}_{\mathcal{F}}$ becomes the residual from the linear regression between X_j and $\mathbf{x}_{\mathcal{F}}$. The normality of $\mathbf{x}$ further implies that

$$\mathrm{var}(X_j \mid \mathbf{x}_{\mathcal{F}}) = \mathrm{var}(X_j) - \mathrm{E}^T(X_j\mathbf{x}_{\mathcal{F}})\boldsymbol{\Sigma}_{\mathcal{F}}^{-1}\mathrm{E}(X_j\mathbf{x}_{\mathcal{F}}) = \mathrm{var}(R_{j|\mathcal{F}}). \tag{2.5}$$

(2.3), (2.4) and (2.5) together guarantee that $\delta_{j|\mathcal{F}} = \mathrm{var}(R_{j|\mathcal{F}})\mathrm{E}\{\partial\Psi_{\mathcal{F},j}(\mathbf{x})/\partial X_j\}$.

For part ii, note that $E(Y \mid \mathbf{x}_{\mathcal{F}}, X_j) = E(Y \mid \mathbf{x}_{\mathcal{F}})$ under H_0. It follows that

$$\mathrm{E}(R_{j|\mathcal{F}}Y) = \mathrm{E}\{R_{j|\mathcal{F}}\mathrm{E}(Y \mid \mathbf{x}_{\mathcal{F}}, X_j)\} = \mathrm{E}\{R_{j|\mathcal{F}}\mathrm{E}(Y \mid \mathbf{x}_{\mathcal{F}})\} = \mathrm{E}\{\mathrm{E}(R_{j|\mathcal{F}} \mid \mathbf{x}_{\mathcal{F}})Y\} = 0.$$

The last equality above holds because $\mathrm{E}(R_{j|\mathcal{F}} \mid \mathbf{x}_{\mathcal{F}}) = \mathrm{E}\{X_j - \mathrm{E}(X_j \mid \mathbf{x}_{\mathcal{F}}) \mid \mathbf{x}_{\mathcal{F}}\} = 0$. □

From Proposition 2.1, we see that a nonzero value of $\delta_{j|\mathcal{F}}$ provides evidence against H_0. In the stepwise regression procedure, we can go over all the indices $j \in \mathcal{F}^c$, select j with a nonzero value of $\delta_{j|\mathcal{F}}$, update the current working index set $\mathcal{F}$ to $\mathcal{F} \cup j$, and then go to the next iteration. If $\delta_{j|\mathcal{F}} = 0$ for all $j \in \mathcal{F}^c$, then we can not reject H_0 for any j, and $\mathcal{F}$ will not be updated. We use this $\mathcal{F}$ to recover the active set $\mathcal{A}_0$. To fix the idea, we consider a toy example as follows.

Example 1. Let $\mathbf{x} = (X_1, X_2, X_3) \sim N(\mathbf{0}, \mathbf{I})$, and $Y = (X_1 + X_2)^3 + \epsilon$, where $\epsilon \sim N(0, 1)$ and $\epsilon \perp\!\!\!\perp \mathbf{x}$. The stepwise regression at the population level then

proceeds as follows. In the initial step, we start from empty working index set $\mathcal{F} = \emptyset$. For $j = 1, 2$, we have $\delta_{j|\emptyset} = 6$. For $j = 3$, we have $\delta_{3|\emptyset} = 0$. By comparing $\delta_{1|\emptyset}$, $\delta_{2|\emptyset}$ and $\delta_{3|\emptyset}$, we either add $j = 1$ or add $j = 2$ into the active set $\mathcal{F}$. Without loss of generality, suppose we add X_1 and the working index set becomes $\mathcal{F} = \{1\}$. In the second step, we have $\delta_{2|\{1\}} = 6$ for $j = 2$ and $\delta_{3|\{1\}} = 0$ for $j = 3$. By comparing $\delta_{j|\mathcal{F}}$ for $j = 2, 3$, we add $j = 2$ into the active set $\mathcal{F}$, and the working index set is now $\mathcal{F} = \{1, 2\}$. In the last step, $\delta_{3|\{1,2\}} = 0$ and $\mathcal{F}$ will not be updated. Thus $\mathcal{F} = \{1, 2\}$ successfully recover the active set $\mathcal{A}_0$.

We see in Proposition 2.1 that $\delta_{j|\mathcal{F}} = 0$ under H_0. If $\delta_{j|\mathcal{F}} = 0$ under H_a as well, then the stepwise regression procedure will no longer work. Conditions (b) and (c) are used to rule out such problematic situations. In the discussions following Proposition 1.3, we have seen that as a dimension reduction method, OLS does not work well when $\mathrm{E}(Y \mid \mathbf{x})$ is symmetric, in which case the stepwise regression procedure will also fail.

Example 2. Let $\mathbf{x} = (X_1, X_2, X_3) \sim N(\mathbf{0}, \mathbf{I})$, and $Y = (X_1 + X_2)^2 + \epsilon$, where $\epsilon \sim N(0, 1)$ and $\epsilon \perp\!\!\!\perp \mathbf{x}$. Start from empty working index set $\mathcal{F} = \emptyset$. We have $\delta_{j|\emptyset} = 0$ for $j = 1, 2, 3$. Thus we can not reject H_0 for any j, and fail to recover $\mathcal{A}_0 = \{1, 2\}$. However, $\partial E(Y \mid \mathbf{x})/\partial X_k = 2(X_1 + X_2)$ for $k = 1, 2$. It follows that $\mathrm{E}\{\partial \mathrm{E}(Y \mid \mathbf{x})/\partial X_k\} = 0$ for $k \in \mathcal{A}_0$ and condition (b) is violated. By assuming condition (b) that $\mathrm{E}\{\partial \mathrm{E}(Y \mid \mathbf{x})/\partial X_k\} \neq 0$ for any $k \in \mathcal{A}_0$, we thus

rule out the case where $\delta_{j|\mathcal{F}} = 0$ under H_a due to the symmetry of $\mathrm{E}(Y \mid \mathbf{x})$, hence the use of condition (b).

Regarding condition (c), in the discussions following Proposition 1.4, we have seen that OLS recovers the central mean space $\mathcal{S}_{\mathrm{E}(Y|\mathbf{x})}$, which can be a proper subspace of the central space $\mathcal{S}_{Y|\mathbf{x}}$. In the case when the conditional distribution $F(Y \mid \mathbf{x})$ depends on the higher-order moments, we may have $\delta_{j|\mathcal{F}} = 0$ under H_a, and the stepwise regression procedure will fail as a result.

Example 3. Let $\mathbf{x} = (X_1, X_2, X_3) \sim N(\mathbf{0}, \mathbf{I})$, and $Y = X_1 + (X_2 + 1)\epsilon$, where $\epsilon \sim N(0, 1)$ and $\epsilon \perp\!\!\!\perp \mathbf{x}$. Start from empty working index set $\mathcal{F} = \emptyset$. We have $\delta_{1|\emptyset} = 1$ and $\delta_{j|\emptyset} = 0$ for $j = 2, 3$. Thus we add $j = 1$ into the active set $\mathcal{F}$. In the next iteration, $\mathcal{F} = \{1\}$ and $\delta_{j|\{1\}} = 0$ for $j = 2, 3$. We can not reject H_0 for $j = 2$ or $j = 3$. Thus $\mathcal{F}$ can not be updated and we stop the iteration. We see that the active set is $\mathcal{A}_0 = \{1, 2\}$ while our final working set is $\mathcal{F} = \{1\}$. Our stepwise regression procedure can effectively recover X_1, which is active due to the regression mean $\mathrm{E}(Y \mid \mathbf{x})$. However, the stepwise regression procedure fails to recover the active predictor X_2, which appears in the variance component. Let $\boldsymbol{\beta}_1 = (1, 0, 0)^T$ and $\boldsymbol{\beta}_2 = (0, 1, 0)^T$. We have $\mathcal{S}_{\mathrm{E}(Y|\mathbf{x})} = \mathrm{span}(\boldsymbol{\beta}_1)$ and $\mathcal{S}_{Y|\mathbf{x}} = \mathrm{span}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$. Thus condition (c) is violated in this example. By assuming condition (c) that $\mathcal{S}_{\mathrm{E}(Y|\mathbf{x})} = \mathcal{S}_{Y|\mathbf{x}}$, we thus rule out the case where $\delta_{j|\mathcal{F}} = 0$ under H_a due to the active predictor in the higher-order moments.

For Example 3 above, we note that $\mathrm{E}\{\partial \mathrm{E}(Y \mid \mathbf{x})/\partial X_1\} = 1$ and $\mathrm{E}\{\partial \mathrm{E}(Y \mid \mathbf{x})/\partial X_2\} = 0$. Thus condition (b) is violated for $k = 2 \in \mathcal{A}_0$. Generally speaking, condition (b) is not always more restrictive than condition (c). For example, consider $Y = X_1 + X_2 + (X_2 + 1)\epsilon$. Then condition (c) is violated but condition (b) holds.

We end this section by revealing the connections between the pivotal quantity $\delta_{j|\mathcal{F}}$ and the OLS estimator $\boldsymbol{\beta}_{\mathrm{OLS}} = \boldsymbol{\Sigma}^{-1}\mathrm{E}(\mathbf{x}Y)$. Let $\boldsymbol{\beta}_{\mathcal{F}} = \boldsymbol{\Sigma}_{\mathcal{F}}^{-1}E(\mathbf{x}_{\mathcal{F}}Y)$ be the $\mathbf{x}$-scale OLS estimator between Y and $\mathbf{x}_{\mathcal{F}}$, and denote $\boldsymbol{\eta}_{\mathcal{F}} = \boldsymbol{\Sigma}_{\mathcal{F}}^{-1/2}E(\mathbf{x}_{\mathcal{F}}Y)$ as the corresponding $\mathbf{z}$-scale OLS estimator. Let $\mathbf{M}_{\mathcal{F}} = \boldsymbol{\Sigma}_{\mathcal{F}}^{-1/2}\mathrm{E}(\mathbf{x}_{\mathcal{F}}Y)\mathrm{E}^T(\mathbf{x}_{\mathcal{F}}Y)\boldsymbol{\Sigma}_{\mathcal{F}}^{-1/2} = \boldsymbol{\eta}_{\mathcal{F}}\boldsymbol{\eta}_{\mathcal{F}}^T$. Similarly we denote $\mathbf{M}_{\mathcal{F}\cup j} = \boldsymbol{\eta}_{\mathcal{F}\cup j}\boldsymbol{\eta}_{\mathcal{F}\cup j}^T$. Denote $\mathrm{tr}(\mathbf{M})$ as the trace of matrix $\mathbf{M}$ and recall that $\delta_{j|\mathcal{F}} = E(YR_{j|\mathcal{F}})$. We have the following result.

Proposition 2.2. *Suppose condition (a) holds. Then* $\mathrm{tr}(\mathbf{M}_{\mathcal{F}\cup j}) - \mathrm{tr}(\mathbf{M}_{\mathcal{F}}) = \{\mathrm{var}(R_{j|\mathcal{F}})\}^{-1}\delta_{j|\mathcal{F}}^2$.

Proof. Let $|\mathcal{F}|$ be the cardinality of $\mathcal{F}$. Denote $\sigma_{j|\mathcal{F}}^2 = \mathrm{var}(R_{j|\mathcal{F}}) = \mathrm{var}(X_j) - \mathrm{E}^T(X_j\mathbf{x}_{\mathcal{F}})\boldsymbol{\Sigma}_{\mathcal{F}}^{-1}\mathrm{E}(X_j\mathbf{x}_{\mathcal{F}})$. Define $(|\mathcal{F}|+1)\times(|\mathcal{F}|+1)$ dimensional matrices $\mathbf{A}$ and $\mathbf{D}$ as

$$\mathbf{A} = \begin{pmatrix} \mathbf{I}_{|\mathcal{F}|} & \mathbf{0} \\ -\mathrm{E}^T(X_j\mathbf{x}_{\mathcal{F}})\boldsymbol{\Sigma}_{\mathcal{F}}^{-1} & 1 \end{pmatrix} \text{ and } \mathbf{D} = \begin{pmatrix} \boldsymbol{\Sigma}_{\mathcal{F}} & \mathbf{0} \\ \mathbf{0} & \sigma_{j|\mathcal{F}}^2 \end{pmatrix}.$$

Then $\mathbf{x}_{\mathcal{F}\cup j} = \{\mathbf{x}_{\mathcal{F}}, X_j\}^T$ and $\mathbf{A}\mathbf{x}_{\mathcal{F}\cup j} = \{\mathbf{x}_{\mathcal{F}}^T, R_{j|\mathcal{F}}\}^T$. Note that $\mathrm{E}(\mathbf{x}_{\mathcal{F}}X_j) = \mathbf{0}$, and we have $\mathrm{var}(\mathbf{A}\mathbf{x}_{\mathcal{F}\cup j}) = \mathbf{D}$. On the other hand, $\mathrm{var}(\mathbf{A}\mathbf{x}_{\mathcal{F}\cup j}) = \mathbf{A}\boldsymbol{\Sigma}_{\mathcal{F}\cup j}\mathbf{A}^T$.

Together we have $\boldsymbol{\Sigma}_{\mathcal{F}\cup j}^{-1} = \mathbf{A}^T\mathbf{D}^{-1}\mathbf{A}$. Note that

$$\begin{aligned}\mathrm{tr}(\mathbf{M}_{\mathcal{F}\cup j}) &= \mathrm{tr}\{\boldsymbol{\Sigma}_{\mathcal{F}\cup j}^{-1}\mathrm{E}(\mathbf{x}_{\mathcal{F}\cup j}Y)\mathrm{E}^T(\mathbf{x}_{\mathcal{F}\cup j}Y)\} = \mathrm{tr}\{\mathbf{D}^{-1}\mathrm{E}(\mathbf{A}\mathbf{x}_{\mathcal{F}\cup j}Y)\mathrm{E}^T(\mathbf{A}\mathbf{x}_{\mathcal{F}\cup j}Y)\} \\ &= \mathrm{tr}\left\{\begin{pmatrix}\boldsymbol{\Sigma}_{\mathcal{F}}^{-1} & \mathbf{0}\\ \mathbf{0} & \sigma_{j|\mathcal{F}}^{-2}\end{pmatrix}\begin{pmatrix}\mathrm{E}(\mathbf{x}_{\mathcal{F}}Y)\\ \mathrm{E}(R_{j|\mathcal{F}}Y)\end{pmatrix}\begin{pmatrix}\mathrm{E}^T(\mathbf{x}_{\mathcal{F}}Y) & \mathrm{E}(R_{j|\mathcal{F}}Y)\end{pmatrix}\right\} \\ &= \mathrm{tr}\{\boldsymbol{\Sigma}_{\mathcal{F}}^{-1}\mathrm{E}(\mathbf{x}_{\mathcal{F}}Y)\mathrm{E}^T(\mathbf{x}_{\mathcal{F}}Y)\} + \sigma_{j|\mathcal{F}}^{-2}\mathrm{E}^2(R_{j|\mathcal{F}}Y) = \mathrm{tr}(\mathbf{M}_{\mathcal{F}}) + \{\mathrm{var}(R_{j|\mathcal{F}})\}^{-1}\delta_{j|\mathcal{F}}^2.\end{aligned}$$

It follows that $\mathrm{tr}(\mathbf{M}_{\mathcal{F}\cup j}) - \mathrm{tr}(\mathbf{M}_{\mathcal{F}}) = \{\mathrm{var}(R_{j|\mathcal{F}})\}^{-1}\delta_{j|\mathcal{F}}^2$. □

Proposition 2.2 is closely related to trace pursuit (Yu, Dong and Zhu, 2015), where slicing-based methods such as SIR, SAVE and directional regression are adapted for model-free variable selection. Our proposal of using $\delta_{j|\mathcal{F}}$ is equivalent to the OLS-based trace pursuit, which does not require slicing of the response Y.

2.2 The Direct Estimation Approach

For the OLS estimator $\boldsymbol{\beta}_{\mathrm{OLS}} = \boldsymbol{\Sigma}^{-1}\mathrm{E}(\mathbf{x}Y) = (\beta_1^{\mathrm{OLS}}, \ldots, \beta_p^{\mathrm{OLS}})^T$ and the full index set $\Omega = \{1, 2, \cdots, p\}$, define

$$\mathcal{A}_{\mathrm{OLS}} = \{k \in \Omega\colon \beta_k^{\mathrm{OLS}} \neq 0\}. \tag{2.6}$$

Proposition 2.3. *Suppose conditions (a) and (b) hold. Then* $\mathcal{A}_{OLS} = \mathcal{A}_0$.

PROOF. Let $\mathbf{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_q)^T$ be the basis of $\mathcal{S}_{Y|\mathbf{x}}$. Then $\text{span}(\mathbf{B}) = \mathcal{S}_{Y|\mathbf{x}}$. From Proposition 1.1, we know that $\boldsymbol{\beta}_{\text{OLS}} \subseteq \mathcal{S}_{Y|\mathbf{x}} = \text{span}(\mathbf{B})$, which means there exist constants $c_1, \ldots, c_q$ such that $\boldsymbol{\beta}_{\text{OLS}} = \sum_{i=1}^{q} c_i \boldsymbol{\beta}_i$. Denote $\boldsymbol{\beta}_i = \{\beta_{i1}, \ldots, \beta_{ip}\}^T$, and we have $\beta_k^{\text{OLS}} = \sum_{i=1}^{q} c_i \beta_{ik}$, $k = 1, \ldots, p$. From the definition of $\mathcal{A}_0$ in (1.3), we know that $\beta_{ik} = 0$ for $i = 1, \ldots, q$ if $k \notin \mathcal{A}_0$. It follows that $\beta_k^{\text{OLS}} = 0$ if $k \notin \mathcal{A}_0$. From the definition of $\mathcal{A}_{\text{OLS}}$ in (2.6), we have $k \notin \mathcal{A}_{\text{OLS}}$ if $k \notin \mathcal{A}_0$. Thus $\mathcal{A}_0^c \subseteq \mathcal{A}_{\text{OLS}}^c$, and $\mathcal{A}_{\text{OLS}} \subseteq \mathcal{A}_0$ as a result.

On the other hand, from Proposition 1.3, we have $\boldsymbol{\beta}_{\text{OLS}} = \text{E}\left\{\frac{\partial \text{E}(Y|\mathbf{x})}{\partial \mathbf{x}}\right\}$. Condition (b) thus guarantees that $\beta_k^{\text{OLS}} \neq 0$ for any $k \in \mathcal{A}_0$. From the definition of $\mathcal{A}_{\text{OLS}}$, we have $k \in \mathcal{A}_{\text{OLS}}$ if $k \in \mathcal{A}_0$. It follows that $\mathcal{A}_0 \subseteq \mathcal{A}_{\text{OLS}}$. □

Proposition 2.3 is a simple yet insightful result. It suggests that we can achieve model-free variable selection by recovering the indices corresponding to the nonzero elements in $\boldsymbol{\beta}_{\text{OLS}}$. To fix the idea, we revisit the models in Examples 1 through 3.

Example 4. Let $\mathbf{x} = (X_1, X_2, X_3) \sim N(\mathbf{0}, \mathbf{I})$. For $Y = (X_1 + X_2)^3 + \epsilon$ in Example 1, $\boldsymbol{\beta}_{\text{OLS}} = (6, 6, 0)^T$ and $\mathcal{A}_{\text{OLS}} = \{1, 2\} = \mathcal{A}_0$. For $Y = (X_1 + X_2)^2 + \epsilon$ in Example 2, $\boldsymbol{\beta}_{\text{OLS}} = (0, 0, 0)^T$. Condition (b) is violated and $\mathcal{A}_{\text{OLS}} = \emptyset \subset \{1, 2\} = \mathcal{A}_0$. For $Y = X_1 + (X_2 + 1)\epsilon$ in Example 3, $\boldsymbol{\beta}_{\text{OLS}} = (1, 0, 0)^T$. Condition (b) is violated and $\mathcal{A}_{\text{OLS}} = \{1\} \subset \{1, 2\} = \mathcal{A}_0$. For $Y = X_1 + X_2 + (X_2 + 1)\epsilon$ in the discussions following Example 3, $\boldsymbol{\beta}_{\text{OLS}} = (1, 1, 0)^T$. Although

condition (c) is violated, condition (b) is satisfied and we have $\mathcal{A}_{\text{OLS}} = \{1, 2\} = \mathcal{A}_0$.

CHAPTER 3

Variable Transformation to Facilitate Model-Free Variable Selection

Model-free variable selection should be flexible so that it can be applied to a broad range of underlying true models. As we have seen in Section 2, the effectiveness of the OLS-based methodology relies on some assumptions. Motivated by recent work in transformed SDR (Wang, 2014; Mai, 2015), we introduce two marginal variable transformation methods, which can be used to expand the capabilities of our proposals in Section 2. We will use variable transformation to address the following common limitations:

L1: Predictor Distribution Assumption. Instead of assuming predictor $\mathbf{x}$ follows the normality assumption, we assume that $\mathbf{x} = (X_1, \ldots, X_p)^T$ can be transformed to $\mathbf{u} = (U_1, \ldots, U_p)^T$ through marginal transformations, and the transformed predictor $\mathbf{u}$ follows the normality assumption.

L2: Symmetric Link Function. We have seen before that condition (b) could fail due to the symmetry of $\mathrm{E}(Y \mid \mathbf{x})$. By marginal transformation of the predictors in an automatic fashion, the potential symmetry in the unknown $\mathrm{E}(Y \mid \mathbf{x})$ will be addressed, and model-free variable selection will become effective again.

L3: Active variables in the Regression Variance Component. We have seen before that condition (c) can be violated due to the active predictors in the higher moments of the conditional distribution $F(Y \mid \mathbf{x})$. We propose response transformation to alleviate this limitation. As an SDR method, OLS, for example, will fail to recover any index solely in the regression variance component. Generally, some SDR methods are able to recover indices in the regression variance component and some are not.

3.1 An Overview of Marginal Predictor Transformation

Consider the mapping $\psi_k \colon \mathbb{R} \mapsto \mathbb{R}$ and the marginal transformations $\psi_k(X_k) = U_k$, $k = 1, \ldots, p$. Then $\mathbf{u} = (U_1, \ldots, U_p)^T$ denotes the marginally transformed predictor. Parallel to (1.3), define index set

$$\mathcal{A} = \{k \in \Omega \colon F(Y \mid \mathbf{u}) \text{ depends on } U_k\}. \tag{3.1}$$

The main idea is to recover the original active predictor set $\mathcal{A}_0$ through estimation of the index set $\mathcal{A}$.

Denote $\mathbf{x}_{\mathcal{A}_0} = \{X_k \colon k \in \mathcal{A}_0\}$ and $\mathbf{u}_{\mathcal{A}} = \{U_k \colon k \in \mathcal{A}\}$. Then the active sets $\mathcal{A}_0$ and $\mathcal{A}$ are the smallest subsets to satisfy

$$Y \perp\!\!\!\perp \mathbf{x} \mid \mathbf{x}_{\mathcal{A}_0} \text{ and } Y \perp\!\!\!\perp \mathbf{u} \mid \mathbf{u}_{\mathcal{A}}. \tag{3.2}$$

Similarly, denote $\mathbf{x}_{\mathcal{A}_0^c} = \{X_k \colon k \in \mathcal{A}_0^c\}$ and $\mathbf{u}_{\mathcal{A}^c} = \{U_k \colon k \in \mathcal{A}^c\}$, where $\mathcal{A}_0^c$ and $\mathcal{A}^c$ are the complements of $\mathcal{A}_0$ and $\mathcal{A}$ respectively. The next result forms the connections between the original active predictor set $\mathcal{A}_0$ and the index set $\mathcal{A}$.

Proposition 3.1. *1. If $\psi_k \colon \mathbb{R} \mapsto \mathbb{R}$ is a one-to-one function for $k = 1, \ldots, p$, then $\mathcal{A} = \mathcal{A}_0$.*

2. If $\mathbf{x}_{\mathcal{A}_0^c} \perp\!\!\!\perp \mathbf{x}_{\mathcal{A}_0}$, then $\mathcal{A} \subseteq \mathcal{A}_0$.

3. *If we have (i)* $(Y, \mathbf{u}_{\mathcal{A}}) \perp\!\!\!\perp \mathbf{u}_{\mathcal{A}^c}$ *guarantees* $(Y, \mathbf{x}_{\mathcal{A}}) \perp\!\!\!\perp \mathbf{x}_{\mathcal{A}^c}$*, and (ii)* $\mathbf{u}_{\mathcal{A}} \perp\!\!\!\perp \mathbf{u}_{\mathcal{A}^c}$*, then* $\mathcal{A}_0 \subseteq \mathcal{A}$.

PROOF. For part 1, because $Y \perp\!\!\!\perp \mathbf{x} \mid \mathbf{x}_{\mathcal{A}_0}$, we have $Y \perp\!\!\!\perp \mathbf{u} \mid \mathbf{x}_{\mathcal{A}_0}$. It follows that $Y \perp\!\!\!\perp \mathbf{u} \mid \mathbf{u}_{\mathcal{A}_0}$ as ψ_k is one-to-one. Note that $\mathcal{A}$ is the smallest set to satisfy $Y \perp\!\!\!\perp \mathbf{u} \mid \mathbf{u}_{\mathcal{A}}$. Thus we have $\mathcal{A} \subseteq \mathcal{A}_0$. It can be shown similarly that $\mathcal{A}_0 \subseteq \mathcal{A}$. Together we have $\mathcal{A} = \mathcal{A}_0$.

For part 2, $\mathbf{x}_{\mathcal{A}_0^c} \perp\!\!\!\perp \mathbf{x}_{\mathcal{A}_0}$ together with $Y \perp\!\!\!\perp \mathbf{x} \mid \mathbf{x}_{\mathcal{A}_0}$ implies that $(Y, \mathbf{x}_{\mathcal{A}_0}) \perp\!\!\!\perp \mathbf{x}_{\mathcal{A}_0^c}$. It follows that $(Y, \mathbf{u}_{\mathcal{A}_0}) \perp\!\!\!\perp \mathbf{u}_{\mathcal{A}_0^c}$, which guarantees that $Y \perp\!\!\!\perp \mathbf{u} \mid \mathbf{u}_{\mathcal{A}_0}$. As a result, we have $\mathcal{A} \subseteq \mathcal{A}_0$.

For part 3, note that $Y \perp\!\!\!\perp \mathbf{u} \mid \mathbf{u}_{\mathcal{A}}$ and assumption (ii) together lead to $(Y, \mathbf{u}_{\mathcal{A}}) \perp\!\!\!\perp \mathbf{u}_{\mathcal{A}^c}$. Assumption (i) then implies that $(Y, \mathbf{x}_{\mathcal{A}}) \perp\!\!\!\perp \mathbf{x}_{\mathcal{A}^c}$ as long as $Y \perp\!\!\!\perp \mathbf{u} \mid \mathbf{u}_{\mathcal{A}}$. As a result, we have $Y \perp\!\!\!\perp \mathbf{x} \mid \mathbf{x}_{\mathcal{A}}$ as long as $Y \perp\!\!\!\perp \mathbf{u} \mid \mathbf{u}_{\mathcal{A}}$. It follows that $\mathcal{A}_0 \subseteq \mathcal{A}$. □

Part 1 of Proposition 3.1 states that if we consider one-to-one marginal transformations, the active set $\mathcal{A}$ based on the transformed predictors coincides with the active set $\mathcal{A}_0$ based on the original predictors. Part 2 and Part 3 provide, respectively, the sufficient conditions for $\mathcal{A}$ to become the subset or the superset of $\mathcal{A}_0$.

3.2 Marginal Predictor Transformation to Achieve Normality

To address limitation 1, Proposition 3.1 suggests that we consider the marginal predictor transformation proposed by Yeo and Johnson (2000). Let $F_k(\cdot)$ be the marginal distribution function of X_k and denote $\Phi^{-1}(\cdot)$ as the inverse of the standard normal distribution function. Consider marginal transformations $\psi_k(\cdot) = \Phi^{-1}(F_k(\cdot))$ and $U_k = \Phi^{-1}\{F_k(X_k)\}$. The Yeo-Johnson transformation has been used in Wang et al. (2014) and Mai and Zou (2015) for transformed SDR and here we establish its use for transformed variable selection. Since the Yeo-Johnson transformation is monotone, we know from Proposition 3.1 that $\mathcal{A} = \mathcal{A}_0$, and the transformation will not affect the recovery of the original active set. Since F_k is unknown in practice, we use the sample distribution function $\hat{F}_k$. For $\mathbf{x}^{(i)} = (X_1^{(i)}, \ldots, X_p^{(i)})^T$, $i = 1, \ldots, n$, we have $\hat{F}_k(X_k^{(i)}) = (n+1)^{-1} r_k^{(i)}$, where $r_k^{(i)} = \sum_{j=1}^{n} I(X_k^{(j)} \leq X_k^{(i)})$ is the rank of $X_k^{(i)}$ among $\{X_k^{(1)}, \ldots, X_k^{(n)}\}$. The sample level transformed predictors become $\hat{\mathbf{u}}^{(i)} = (\hat{U}_1^{(i)}, \ldots, \hat{U}_p^{(i)})^T$, where $\hat{U}_k^{(i)} = \Phi^{-1}\{(n+1)^{-1} r_k^{(i)}\}$ for $i = 1, \ldots, n$ and $k = 1, \ldots, p$.

The purpose of the Yeo-Johnson transformation is to achieve normality of the transformed predictor. Note that instead of assuming the normality of $\mathbf{x}$, we now assume that $\mathbf{u} = (U_1, \ldots, U_p)^T$ is multivariate normal with mean

$\mathbf{0}$ and variance $\boldsymbol{\Sigma}$ after the Yeo-Johnson transformation. This assumption was made previously by Mai and Zou (2015). Proposition 1 of Mai and Zou (2015) states that if there exists a set of increasing univariate functions $\boldsymbol{\Psi} = (\psi_1, \ldots, \psi_p)$ such that $\boldsymbol{\Psi}(\mathbf{x}) = \{\psi_1(X_1), \ldots, \psi_p(X_p)\}^T$ is $N(\mathbf{0}, \boldsymbol{\Sigma})$, then we must have $\psi_k(\cdot) = \Phi^{-1}(F_k(\cdot))$ for $k = 1, \ldots, p$.

3.3 Marginal Predictor Transformation to Address Symmetric Link Functions

For the ease of presentation, we focus on the direct estimation approach in this section. Denote $\boldsymbol{\Sigma}^{\mathbf{u}} = \mathrm{var}(\mathbf{u})$ and $\boldsymbol{\beta}^{\mathbf{u}}_{\mathrm{OLS}} = (\boldsymbol{\Sigma}^{\mathbf{u}})^{-1}\mathrm{E}(\mathbf{u}Y) = (\beta^{\mathbf{u}}_1, \ldots, \beta^{\mathbf{u}}_p)^T$. Similar to the definition of $\mathcal{A}_{\mathrm{OLS}}$ in (2.6), define

$$\mathcal{A}^{\mathbf{u}}_{\mathrm{OLS}} = \{k \in \Omega \colon \beta^{\mathbf{u}}_k \neq 0\}. \tag{3.3}$$

Recall that $\mathcal{A}_0 = \{k \in \Omega \colon F(Y \mid \mathbf{x})$ depends on $X_k\}$. In the case when $\mathrm{E}(Y \mid \mathbf{x})$ is symmetric for X_k with $k \in \mathcal{A}_0$, we have seen that $k \notin \mathcal{A}_{\mathrm{OLS}}$, and $\mathcal{A}_{\mathrm{OLS}}$ becomes a proper subset of $\mathcal{A}_0$. Our idea is to apply marginal predictor transformation $U_k = \psi_k(X_k)$, such that $\mathcal{A}^{\mathbf{u}}_{\mathrm{OLS}}$ can be used to recover $\mathcal{A}_0$. We use the following example to fix the idea.

Example 5. Let $\mathbf{x} = (X_1, X_2, X_3)^T \sim N(\mathbf{0_3}, \mathbf{I_3})$, $\epsilon \sim N(0, 1)$, $\epsilon \perp\!\!\!\perp \mathbf{x}$, and $Y = X_1^2 + X_2 + \epsilon$. Then $\boldsymbol{\beta}_{\mathrm{OLS}} = \boldsymbol{\Sigma}^{-1}\mathrm{E}(\mathbf{x}Y) = (0, 1, 0)^T$, and $\mathcal{A}_{\mathrm{OLS}} = \{2\} \subset \{1, 2\} = \mathcal{A}_0$. Consider $\mathbf{u} = (U_1, U_2, U_3)$, where $U_1 = X_1^2 - \mathrm{E}(X_1^2)$, $U_2 = X_2$

and $U_3 = X_3$. Then $\mathbf{\Sigma^u} = \text{var}(\mathbf{u}) = \text{diag}(2,1,1)$, $\text{E}(\mathbf{u}Y) = (2,1,0)^T$, and $\boldsymbol{\beta}^{\mathbf{u}}_{\text{OLS}} = (\mathbf{\Sigma^u})^{-1}\text{E}(\mathbf{u}Y) = (1,1,0)^T$. From the definition of $\mathcal{A}^{\mathbf{u}}_{\text{OLS}}$ in (3.3), it follows that $\mathcal{A}^{\mathbf{u}}_{\text{OLS}} = \{1,2\} = \mathcal{A}_0$.

The challenge with the approach above is how to determine the correct transformations for different components of $\mathbf{u}$. In Example 5, suppose we use the square transformation to all the components of $\mathbf{x}$, such that $\mathbf{u} = (U_1, U_2, U_3)$, where $U_k = X_k^2 - \text{E}(X_k^2)$, for $k = 1,2,3$. Then $\mathbf{\Sigma^u} = \text{var}(\mathbf{u}) = \text{diag}(2,2,2)$, $\text{E}(\mathbf{u}Y) = (2,0,0)^T$, and $\boldsymbol{\beta}^{\mathbf{u}}_{\text{OLS}} = (\mathbf{\Sigma^u})^{-1}\text{E}(\mathbf{u}Y) = (1,0,0)^T$. It follows that $\mathcal{A}^{\mathbf{u}}_{\text{OLS}} = \{1\} \subset \{1,2\} = \mathcal{A}_0$. Because $Y = X_1^2 + X_2 + \epsilon$ is quadratic in X_1 and linear in X_2, the square transformation of X_1 works well and the square transformation of X_2 does not work.

Next we propose an automatic procedure to decide the proper transformations for different components of $\mathbf{u}$. Denote $\mathbf{x}_{(k)} = (\tilde{X}_k, \tilde{X}_k^2, \ldots, \tilde{X}_k^M)^T$, where $\tilde{X}_k^m = X_k^m - \text{E}(X_k^m)$, $m = 1, \ldots, M$. If $\text{E}(Y\mathbf{x}_{(k)}) = \mathbf{0}$, then set $U_k = X_k$. If $\text{E}(Y\mathbf{x}_{(k)}) \neq \mathbf{0}$, set the marginal transformation of X_k as $U_k = \text{E}^T(Y\mathbf{x}_{(k)})\mathbf{\Sigma}^{-1}_{(k)}\mathbf{x}_{(k)}$, where $\mathbf{\Sigma}_{(k)} = \text{var}(\mathbf{x}_{(k)})$. We revisit Example 5 to fix the idea.

Example 5. (Continued) Let $\mathbf{x} = (X_1, X_2, X_3)^T \sim N(\mathbf{0_3}, \mathbf{I_3})$, $\epsilon \sim N(0,1)$, $\epsilon \perp\!\!\!\perp \mathbf{x}$, and $Y = X_1^2 + X_2 + \epsilon$. Set $M = 2$. Then for $k = 1,2,3$, $\mathbf{x}_{(k)} = \{X_k, X_k^2 -$

$\mathrm{E}(X_k^2)\}^T$ and $\mathbf{\Sigma}_{(k)} = \mathrm{var}(\mathbf{x}_{(k)}) = \mathrm{diag}(1, 2)$. Since $\mathrm{E}(Y\mathbf{x}_{(1)}) = (0, 2)^T$, we have

$$U_1 = \mathrm{E}^T(Y\mathbf{x}_{(1)})\mathbf{\Sigma}_{(1)}^{-1}\mathbf{x}_{(1)} = (0, 2)\begin{pmatrix} 1 & 0 \\ 0 & 1/2 \end{pmatrix}\begin{pmatrix} X_1 \\ X_1^2 - \mathrm{E}(X_1^2) \end{pmatrix} = X_1^2 - \mathrm{E}(X_1^2).$$

Since $\mathrm{E}(Y\mathbf{x}_{(2)}) = (1, 0)^T$, we have

$$U_2 = \mathrm{E}^T(Y\mathbf{x}_{(2)})\mathbf{\Sigma}_{(2)}^{-1}\mathbf{x}_{(2)} = (1, 0)\begin{pmatrix} 1 & 0 \\ 0 & 1/2 \end{pmatrix}\begin{pmatrix} X_2 \\ X_2^2 - \mathrm{E}(X_2^2) \end{pmatrix} = X_2.$$

Since $\mathrm{E}(Y\mathbf{x}_{(3)}) = (0, 0)^T$, we have $U_3 = X_3$. Thus we have automatically choose the square transformation for X_1 while keep X_2 and X_3 the same.

We describe the sample level data-driven transformation as follows. Let $\{(\mathbf{x}^{(i)}, Y^{(i)})\colon i = 1, \ldots, n\}$ be an i.i.d. sample, where $\mathbf{x}^{(i)} = (X_1^{(i)}, \ldots, X_p^{(i)})^T$. For $k = 1, \ldots, p$, we considering fitting Y through M^{th} order polynomial of X_k for some fixed integer M, and the fitted response is used as U_k. Specifically, let $\hat{\mathbf{u}}_k = (\hat{U}_k^{(1)}, \ldots, \hat{U}_k^{(n)}) \in \mathbb{R}^n$ be the collection of transformed predictor U_k at the sample level, $\mathbf{y} = (Y^{(1)}, \ldots, Y^{(n)})^T \in \mathbb{R}^n$ and $\mathbf{\Xi}_k \in \mathbb{R}^{n\times(M+1)}$, where the ith row of $\mathbf{\Xi}_k$ is $\{1, X_k^{(i)}, (X_k^{(i)})^2, \ldots, (X_k^{(i)})^M\}$ for $i = 1, \ldots, n$. Then we have $\hat{\mathbf{u}}_k = \mathbf{\Xi}_k(\mathbf{\Xi}_k^T\mathbf{\Xi}_k)^{-1}\mathbf{\Xi}_k^T\mathbf{y}$.

3.4 Response Transformation to Address Active Variables in the Regression Variance Component

For the ease of presentation, we focus on the direct estimation approach in this section. As we have seen in Section 1.3, if the conditional distribution of Y given $\mathbf{x}$ depends only on up to the Kth moments of $\mathbf{x}$, then $\mathcal{S}_{Y|\mathbf{x}} = \mathcal{S}^{(K)}_{Y|\mathbf{x}}$. We focus on the case when $F(Y \mid \mathbf{x})$ depends on $\mathbf{x}$ through the regression mean $\mathrm{E}(Y \mid \mathbf{x})$ and the regression variance $\mathrm{var}(Y \mid \mathbf{x})$. Equivalently, $\mathcal{S}_{Y|\mathbf{x}} = \mathcal{S}^{(2)}_{Y|\mathbf{x}}$. From Proposition 1.4, we know $\boldsymbol{\Sigma}^{-1}\mathrm{E}(\mathbf{x}Y^k) \subseteq \mathcal{S}^{(2)}_{Y|\mathbf{x}} = \mathcal{S}_{Y|\mathbf{x}}$ for $k = 1, 2$ in this case. Denote $\boldsymbol{\beta}^{Y2}_{\mathrm{OLS}} = \boldsymbol{\Sigma}^{-1}\mathrm{E}(\mathbf{x}Y^2) = (\beta^{Y2}_1, \ldots, \beta^{Y2}_p)^T$. Similar to the definition of $\mathcal{A}_{\mathrm{OLS}}$ in (2.6), define

$$\mathcal{A}^{Y2}_{\mathrm{OLS}} = \{k \in \Omega \colon \beta^{Y2}_k \neq 0\}. \tag{3.4}$$

Then we can recover the active predictor set $\mathcal{A}_0$ by the union of $\mathcal{A}_{\mathrm{OLS}}$ and $\mathcal{A}^{Y2}_{\mathrm{OLS}}$ in (3.4).

To fix the idea, we revisit Example 3 from Section 2.1.

Example 3 (Continued). Let $\mathbf{x} = (X_1, X_2, X_3) \sim N(\mathbf{0}, \mathbf{I})$, and $Y = X_1 + (X_2 + 1)\epsilon$, where $\epsilon \sim N(0, 1)$ and $\epsilon \perp\!\!\!\perp \mathbf{x}$. Since $\boldsymbol{\beta}_{\mathrm{OLS}} = \boldsymbol{\Sigma}^{-1}\mathrm{E}(\mathbf{x}Y) = (1, 0, 0)^T$, $\mathcal{A}_{\mathrm{OLS}} = \{1\}$. Since $\boldsymbol{\beta}^{Y2}_{\mathrm{OLS}} = \boldsymbol{\Sigma}^{-1}\mathrm{E}(\mathbf{x}Y^2) = (0, 2, 0)^T$, $\mathcal{A}^{Y2}_{\mathrm{OLS}} = \{2\}$. Then $\mathcal{A}_{\mathrm{OLS}} \cup \mathcal{A}^{Y2}_{\mathrm{OLS}} = \{1, 2\} = \mathcal{A}_0$.

From the example above, we see that $\boldsymbol{\beta}_{\text{OLS}}$ can help recover active predictors in the regression mean component, while $\boldsymbol{\beta}^{Y2}_{\text{OLS}}$, the OLS estimator based on the transformed response, can help recover active predictors in the regression variance component. By taking the union of $\mathcal{A}_{\text{OLS}}$ and $\mathcal{A}^{Y2}_{\text{OLS}}$, we recover the active predictors in either the regression mean or the regression variance. In the case when $\mathcal{S}_{Y|\mathbf{x}} = \mathcal{S}^{(K)}_{Y|\mathbf{x}}$ for a general K, we can define $\mathcal{A}^{Yk}_{\text{OLS}}$ as the indices corresponding to the nonzero elements of $\boldsymbol{\beta}^{Yk}_{\text{OLS}} = \boldsymbol{\Sigma}^{-1}\text{E}(\mathbf{x}Y^k)$, $k = 1, \ldots, K$. Then we recover $\mathcal{A}_0$ by the union $\cup_{k=1}^{K}\mathcal{A}^{Yk}_{\text{OLS}}$, where $\mathcal{A}^{Y1}_{\text{OLS}} = \mathcal{A}_{\text{OLS}}$.

CHAPTER 4

Sample Level Algorithms for Model-Free Variable Selection

Suppose $\{(\mathbf{x}^{(i)}, Y^{(i)}) : i = 1, \ldots, n\}$ is an i.i.d. sample with $\mathbf{x}^{(i)} = (X_1^{(i)}, \ldots, X_p^{(i)})^T$. We provide two algorithms to recover the active predictor set $\mathcal{A}_0$ in this section. In the cases when one or more predictor transformations are needed, (See Section 3.2 and Section 3.3), we apply the proposed algorithms to $(\hat{\mathbf{u}}^{(i)}, Y^{(i)})$ with $\hat{\mathbf{u}}^{(i)} = (\hat{U}_1^{(i)}, \ldots, \hat{U}_p^{(i)})^T$. In the case when response transformation is needed (See Section 3.4), we apply the proposed algorithms to $\{\mathbf{x}^{(i)}, (Y^{(i)})^2\}$. In the case when both predictor transformation and response transformation are needed, we apply the proposed algorithms to $\{\hat{\mathbf{u}}^{(i)}, (Y^{(i)})^2\}$. Without loss of generality, the proposed algorithms are presented with the untransformed observations $(\mathbf{x}^{(i)}, Y^{(i)})$.

4.1 A Permutation Test Approach and the Stepwise Regression Algorithm

First we restate the stepwise regression hypotheses introduced in Section 2.1 as follows:

$$H_0 \colon Y \perp\!\!\!\perp X_j \mid \mathbf{x}_{\mathcal{F}} \text{ v.s. } H_a \colon Y \text{ is not independent of } X_j \text{ given } \mathbf{x}_{\mathcal{F}}.$$

We have seen in Proposition 2.1 that the pivotal quantity $\delta_{j|\mathcal{F}} = \mathrm{E}(YR_{j|\mathcal{F}})$ with $R_{j|\mathcal{F}} = X_j - \mathrm{E}(X_j \mid \mathbf{x}_{\mathcal{F}})$ plays an important role in testing the stepwise regression hypotheses.

Assume without loss of generality that $\sum_{i=1}^n Y^{(i)} = 0$ and $\sum_{i=1}^n \mathbf{x}^{(i)} = \mathbf{0}$. For any index set $\mathcal{F}$, denote $\mathbf{x}_{\mathcal{F}}^{(i)} = \{X_k^{(i)} : k \in \mathcal{F}\}$. Let $\hat{\sigma}_j^2 = n^{-1}\sum_{i=1}^n (X_j^{(i)})^2$, $\hat{\mathbf{\Sigma}}_{\mathcal{F}} = n^{-1}\sum_{i=1}^n \mathbf{x}_{\mathcal{F}}^{(i)}(\mathbf{x}_{\mathcal{F}}^{(i)})^T$ and $\hat{\mathbf{\Sigma}}_{\mathcal{F},j} = n^{-1}\sum_{i=1}^n \mathbf{x}_{\mathcal{F}}^{(i)} X_j^{(i)}$. It follows that $\hat{R}_{j|\mathcal{F}}^{(i)} = X_j^{(i)} - \hat{\mathbf{\Sigma}}_{\mathcal{F},j}^T \hat{\mathbf{\Sigma}}_{\mathcal{F}}^{-1} \mathbf{x}_{\mathcal{F}}^{(i)}$ and $\hat{\sigma}_{j|\mathcal{F}}^2 = \hat{\sigma}_j^2 - \hat{\mathbf{\Sigma}}_{\mathcal{F},j}^T \hat{\mathbf{\Sigma}}_{\mathcal{F}}^{-1} \hat{\mathbf{\Sigma}}_{\mathcal{F},j}$. The sample level test statistic for the stepwise regression hypotheses is defined as

$$T_{j|\mathcal{F}} = n\hat{\sigma}_{j|\mathcal{F}}^{-2} \left(n^{-1} \sum_{i=1}^n \hat{R}_{j|\mathcal{F}}^{(i)} Y^{(i)} \right)^2. \tag{4.1}$$

Before we use $T_{j|\mathcal{F}}$ to formally test the stepwise regression hypotheses, we need to find the asymptotic distribution of $T_{j|\mathcal{F}}$ under H_0. We make the following key observation.

Proposition 4.1. *Suppose* $\mathbf{x}$ *is normal with* $E(\mathbf{x}) = \mathbf{0}$. *Then* $R_{j|\mathcal{F}} \perp\!\!\!\perp (\mathbf{x}_{\mathcal{F}}, Y)$ *under* $H_0 : Y \perp\!\!\!\perp X_j \mid \mathbf{x}_{\mathcal{F}}$.

PROOF. From $R_{j|\mathcal{F}} = X_j - \mathrm{E}(X_j \mid \mathbf{x}_\mathcal{F})$ and $Y \perp\!\!\!\perp X_j \mid \mathbf{x}_\mathcal{F}$, we have $Y \perp\!\!\!\perp (X_j, \mathbf{x}_\mathcal{F}) \mid \mathbf{x}_\mathcal{F}$. Thus $Y \perp\!\!\!\perp R_{j|\mathcal{F}} \mid \mathbf{x}_\mathcal{F}$ under H_0. The normality assumption implies that $R_{j|\mathcal{F}} \perp\!\!\!\perp \mathbf{x}_\mathcal{F}$. Together we have $R_{j|\mathcal{F}} \perp\!\!\!\perp (\mathbf{x}_\mathcal{F}, Y)$. □

Proposition 4.1 implies the asymptotic null distribution of $T_{j|\mathcal{F}}$ can be approximated through random permutations of $\{Y^{(i)} : i = 1, \ldots, n\}$. More specifically, Proposition 4.1 motivates us to consider the following permutation test procedure.

P1. Based on $\{(\mathbf{x}^{(i)}, Y^{(i)}) : i = 1, \ldots, n\}$, calculate $T_{j|\mathcal{F}}$ as in (4.1).

P2. Fix $\{\mathbf{x}^{(i)} : i = 1, \ldots, n\}$. For $b = 1, \ldots, B$, denote $\{Y^{(i)}_{\{b\}} : i = 1, \ldots, n\}$ as the b^{th} random permutation of $\{Y^{(i)} : i = 1, \ldots, n\}$. Then calculate $T^{\{b\}}_{j|\mathcal{F}}$ based on the permuted sample $\{(\mathbf{x}^{(i)}, Y^{(i)}_{\{b\}}) : i = 1, \ldots, n\}$.

P3. Calculate the approximate p-value as $p_{j|\mathcal{F}} = B^{-1} \sum_{b=1}^{B} I(T^{\{b\}}_{j|\mathcal{F}} > T_{j|\mathcal{F}})$, where $I(\cdot)$ denotes the indicator function. For a given significance level α, reject $H_0 : Y \perp\!\!\!\perp X_j \mid \mathbf{x}_\mathcal{F}$ if $p_{j|\mathcal{F}} < \alpha$.

The permutation test procedure above is quite straightforward to implement; it can applied to any response transformation, and has the additional benefit of not requiring the exact asymptotic expansion of $T_{j|\mathcal{F}}$.

We now describe the stepwise regression algorithm to recover the active set $\mathcal{A}_0$. The algorithm is parallel to the classical stepwise regression in linear

models, and the main difference is that our method uses the p-value from the permutation test procedure described above.

S1. Set initial working set $\mathcal{F} = \emptyset$. Let p_a and p_d be presepcified thresholds for the addition step and the deletion step respectively.

S2. Addition step: find index j_a such that $j_a = \arg\min_{j \in \mathcal{F}^c} p_{j|\mathcal{F}}$. If $p_{j|\mathcal{F}} < p_a$, update $\mathcal{F}$ to be $\mathcal{F} \cup j_a$. If $p_{j|\mathcal{F}} \geq p_a$, no predictor is added.

S3. Deletion step: find index j_d such that $j_d = \arg\max_{j \in \mathcal{F}} p_{j|(\mathcal{F} \cap \{j\}^c)}$. If $p_{j_d|(\mathcal{F} \cap \{j_d\}^c)} > p_d$, update $\mathcal{F}$ to be $\mathcal{F} \cap \{j_d\}^c$. If $p_{j_d|(\mathcal{F} \cap \{j_d\}^c)} \leq p_d$, no predictor is deleted.

S4. Iterate between steps 2 and 3 until no predictors can be added or deleted.

We remark that the stepwise regression algorithm above is applicable in the $p > n$ setting. Since the algorithm involves inversion of $|\mathcal{F}| \times |\mathcal{F}|$-dimensional matrix $\hat{\mathbf{\Sigma}}_{\mathcal{F}}$, it is computationally feasible as long as $|\mathcal{F}| < p$.

4.2 The Direct Estimation Algorithm through L1 Penalized Regression

Recall from Section 2.2 that $\mathcal{A}_{\text{OLS}} = \{k \in \Omega \colon \beta_k^{\text{OLS}} \neq 0\}$, and that the direct estimation approach recovers the active predictor set $\mathcal{A}_0$ through $\mathcal{A}_{\text{OLS}}$

at the population level. Successful implementation of the direct estimation approach relies on estimating some elements of $\boldsymbol{\beta}_{\text{OLS}} = (\beta_1^{\text{OLS}}, \ldots, \beta_p^{\text{OLS}})^T$ to be exactly zero at the sample level. Unfortunately, if elements of $\boldsymbol{\beta}_{\text{OLS}}$ are close yet not exactly zero, we cannot use the classical hypothesis tests as the violation of the linear model assumption may lead to inaccurate hypothesis test results.

Assume without loss of generality that $\sum_{i=1}^n Y^{(i)} = 0$ and $\sum_{i=1}^n \mathbf{x}^{(i)} = \mathbf{0}$. The sample level OLS estimator is $\hat{\boldsymbol{\beta}}_{\text{OLS}} = n^{-1}\hat{\boldsymbol{\Sigma}}^{-1}\sum_{i=1}^n \mathbf{x}^{(i)}Y^{(i)}$, where $\hat{\boldsymbol{\Sigma}} = n^{-1}\sum_{i=1}^n \mathbf{x}^{(i)}(\mathbf{x}^{(i)})^T$. The OLS estimator can be viewed as the minimizer of sums of least squares

$$\hat{\boldsymbol{\beta}}_{\text{OLS}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \sum_{i=1}^n (Y^{(i)} - \boldsymbol{\beta}^T\mathbf{x}^{(i)})^2.$$

Let $\hat{\boldsymbol{\beta}}_{\text{OLS}} = (\hat{\beta}_1^{\text{OLS}}, \ldots, \hat{\beta}_p^{\text{OLS}})^T$ and denote $\hat{\mathcal{A}}_0^{\text{OLS}} = \{k \in \Omega\colon \hat{\beta}_k^{\text{OLS}} \neq 0\}$. Since $\hat{\boldsymbol{\beta}}_{\text{OLS}}$ is not a sparse estimator of $\boldsymbol{\beta}_{\text{OLS}}$, it turns out that $\hat{\mathcal{A}}_0^{\text{OLS}} = \Omega$ is a bad estimator of $\mathcal{A}_0$.

To encourage sparsity, LASSO (Tibshirani, 1996) considers $L1$ penalized regression and estimate $\boldsymbol{\beta}_{\text{OLS}}$ by

$$\hat{\boldsymbol{\beta}}_{\text{LASSO}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \sum_{i=1}^n (Y^{(i)} - \boldsymbol{\beta}^T\mathbf{x}^{(i)})^2 \text{ subject to } \|\boldsymbol{\beta}\|_1 < \lambda.$$

Here $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^p |\beta_j|$ is the $L1$ norm of $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$, and λ is a tuning parameter. Larger values of λ will result in fewer coefficients estimated as zero, and a smaller λ value will lead to more zero coefficients. Let $\hat{\boldsymbol{\beta}}_{\text{LASSO}} =$

$(\hat{\beta}_1^{\text{LASSO}}, \ldots, \hat{\beta}_p^{\text{LASSO}})^T$ and denote $\hat{\mathcal{A}}_0^{\text{LASSO}} = \{k \in \Omega\colon \hat{\beta}_k^{\text{LASSO}} \neq 0\}$. The active predictor set $\mathcal{A}_0$ can then be estimated by $\hat{\mathcal{A}}_0^{\text{LASSO}}$.

The Dantzig selector (Candes and Tao, 2007) is another popular $L1$ penalized regression approach. The Dantzig selector estimates $\boldsymbol{\beta}_{\text{OLS}}$ by the solution of

$$\min \|\boldsymbol{\beta}\|_1, \text{ such that } \|\mathbf{x}^T(Y - \boldsymbol{\beta}^T\mathbf{x})\|_\infty < \omega. \tag{4.2}$$

Here $\|\cdot\|_\infty$ denotes the L-infinity norm. For $\mathbf{a} = (a_1, \ldots, a_p)$, its L-infinity norm is $\|\mathbf{a}\|_\infty = \max\{|a_1|, \ldots, |a_p|\}$. Let $\hat{\boldsymbol{\beta}}_{\text{Dantzig}} = (\hat{\beta}_1^{\text{Dantzig}}, \ldots, \hat{\beta}_p^{\text{Dantzig}})^T$ be the solution of (4.2). Denote $\hat{\mathcal{A}}_0^{\text{Dantzig}} = \{k \in \Omega\colon \hat{\beta}_k^{\text{Dantzig}} \neq 0\}$. The active predictor set $\mathcal{A}_0$ can then be estimated by $\hat{\mathcal{A}}_0^{\text{Dantzig}}$.

We remark that the LASSO estimator $\hat{\boldsymbol{\beta}}_{\text{LASSO}}$ can be viewed as the solution of

$$\min \|\boldsymbol{\beta}\|_1, \text{ such that } \|Y - \boldsymbol{\beta}^T\mathbf{x}\|_2^2 < \lambda^*. \tag{4.3}$$

Here $\|\cdot\|_2$ denotes the L-2 norm. For $\mathbf{a} = (a_1, \ldots, a_p)$, its L-2 norm is $\|\mathbf{a}\|_2 = \sqrt{\sum_{j=1}^p a_j^2}$. Due to the similar forms of (4.2) and (4.3), it turns out that $\hat{\boldsymbol{\beta}}_{\text{LASSO}}$ and $\hat{\boldsymbol{\beta}}_{\text{Dantzig}}$ are quite close to each other in the linear regression setting. Efron et al. (2007) outlined the similarities and dissimilarities between the two and James et al. (2009) further examined the similarities between the two and studied the exact conditions for them to produce identical results. Without assuming the linear regression model, numerical studies are presented

in Section 5 to compare the performances of $\hat{\boldsymbol{\beta}}_{\text{LASSO}}$ and $\hat{\boldsymbol{\beta}}_{\text{Dantzig}}$ for model-free variable selection.

In the procedures described above, $\hat{\mathcal{A}}_0^{\text{LASSO}}$ implicitly depends on the tuning parameter λ, and $\hat{\mathcal{A}}_0^{\text{Dantzig}}$ implicitly depends on the tuning parameter ω. Existing tuning parameter selection methods for LASSO and Dantzig selector relies on the linear model assumption, and can not be used directly in our setting. We use Dantzig selector as an example and describe how to select ω. Similar procedures can be used to select λ for LASSO. Let $\omega_1, \ldots, \omega_H$ be a set of ω values to choose from. Without loss of generality, suppose $\hat{\mathcal{A}}_0^{\text{Dantzig}}$ always contains the first $c(\omega_h)$ predictors $X_1, \ldots, X_{c(\omega_h)}$ when we set $\omega = \omega_h$, $h = 1, \ldots, H$. For unknown link function between Y and $\mathbf{x}$, we fit the following additive model

$$Y = g_1(X_1) + \cdots + g_{c(\omega_h)}(X_{c(\omega_h)}) + \epsilon$$

and get $\hat{g}_1, \ldots, \hat{g}_{c(\omega_h)}$. The fitted response is then $\hat{Y}^{(i)}_{c(\omega_h)} = \hat{g}_1(X_1^{(i)}) + \cdots + \hat{g}_{c(\omega_h)}(X^{(i)}_{c(\omega_h)})$. The sums of squares of error is thus $SSE(\omega_h) = \sum_{i=1}^n (\hat{Y}^{(i)}_{c(\omega_h)} - Y^{(i)})^2$. The ω is then chosen such that $SSE(\omega)$ is minimized among $\omega = \omega_h$, $h = 1, \ldots, H$. Instead of the in-sample sums of squares of error, cross-validation may be used to compute the out-of-sample sums of squares of error, which can then be used for selecting the optimal ω in (4.2). The choice of the $g(\cdot)$ functions in the additive model is flexible, and we use smoothing splines in our numerical studies in the next section.

CHAPTER 5

Numerical Studies

5.1 Simulation Studies

Now we want to examine the performance the algorithms described in the previous chapter through extensive simulation studies. For reference purposes we refer to the stepwise procedure described in Section 4.1 as MFVS-S and we'll refer to the penalized version describe in Section 4.2 as MFVS-D, where the 'D' indicates the Dantzig selector. For comparison purposes we also look at the performance of LASSO with the standard linear cross validation procedure and another SDR-based model free variable selection procedure called COP (Zhong et al., 2012).

Our goal is to understand how our proposed sample level algorithms perform with and without the transformations described in Section 3 using data simulated from six different models:

1. Linear:

   $Y = \boldsymbol{\beta}^T \mathbf{x} + \varepsilon$, where $\boldsymbol{\beta} = (0.6, 1.4, -2, -0.5)$

2. NonLinear Link Function:

   $Y = X_2^3 + \exp(X_2) + (X_3 + X_4)^{1.5} + \varepsilon$

3. Non-Linear Link Function + Non-Normal Predictors:

   $Y = X_2^3 + \exp(X_2) + (X_3 + X_4)^{1.5} + \varepsilon$

4. Heteroscedastic Error:

   $Y = X_1 + X_2 + X_3 + X_4\varepsilon$

5. Quadratic Link Function:

   $Y = X_1 + 0.75X_2^2 + X_3 + 0.5X_4^4 + \varepsilon$

6. Combination of Models 2 to 5:

   $Y = X_1^2 + X_2 + X_3^4 + X_4\varepsilon$

For the data simulated for models 1, 2, 4 and 5, $\mathbf{x}$ has a multivariate normal distribution with mean $\mathbf{0}$ and Toeplitz covariance matrix $\boldsymbol{\Sigma} = \{\sigma_{ij}\}$, where $\sigma_{ij} = \rho^{|i-j|}$. For model 3, $\mathbf{x}$ has a skewed Laplace distribution with parameters $(2, 6, 0)$ and the same Toeplitz covariance structure as the other models. For

model 6, X_2 has the distribution used in model 2 and the other predictors are, again, multivariate normal. For all five models, the error term has a normal distribution with mean zero and variance σ_ε.

We also simulate data for each model across a variety of predictor sizes and across varying degrees of predictor correlation by varying $p \in \{10, 50, 100, 200\}$ and $\rho \in \{0, 0.1, \cdots, 0.9\}$. The combinations of ρ and p lead to 40 versions of each model to simulate and we will simulate each of the 40 scenarios 100 times. Note that, by design, in each simulated dataset $\mathcal{A} = (1, 2, 3, 4)$.

For each of the 40 sets of conditions, we'll select variables using MFVS-S, MFVS-D, LASSO and COP. For MFVS-S and MFVS-D we'll use 5 versions of each, one with all transformations, 3 with each of the 3 transformations only and one with no transformation. For MFVS-S we use the permutation test described in Section 4.1 and for MFVS-D we use cross validation as described in Section 4.2.

Our overall metrics of accuracy will be false positive rate and false negative rate. False positive rate here is defined as the proportion of variables that are not in the true active set yet are selected by the method:

$$\text{FPR} = \frac{|\mathcal{F} \cap \mathcal{A}^c|}{p - |\mathcal{A}|}.$$

False negative rate is defined as the proportion of variables that are in the true active set but are not selected by the method:

$$\text{FNR} = 1 - \frac{|\mathcal{A} \cap \mathcal{F}|}{|\mathcal{A}|}.$$

The figures are constructed to show false positive rates and false negative rates only for each method and model. The tables show, in addition to false positive and false negative rates, the frequency at which each of the four predictors in $\mathcal{A}$ were correctly selected by each method. The tables also show the size of the set of selected variables. Since each of the simulated datasets has 4 active predictors, the frequency of selection is indicated as f_i, where $i = (1, 2, 3, 4)$, indicating each of the four active predictors. The size of the selected set of variables is indicated in the table as $|\mathcal{F}|$. Note that COP is not able to perform variable selection when $p > n$ and this is reflected in the tables and figures.

These extensive simulations show that MFVS-S and MFVS-D work quite will in the nonlinear setting and that the proposed transformations are very effective. We will discuss the simulation results pertaining to each of the six models in the next six sections.

5.1.1 Model 1: Linear

The results for model 1 are shown in Tables 5.1 and 5.2 and in Figures 5.1, 5.2, 5.3, 5.4 and 5.5. Recall that model one is:

$$Y = \boldsymbol{\beta}^T \mathbf{x} + \varepsilon, \text{ where } \boldsymbol{\beta} = (0.6, 1.4, -2, -0.5).$$

Since model 1 is a linear model, we expect all all four methods to do quite well, i.e. to have near zero false positive and false negative rates. This is true for all of the methods but COP. While it is true that COP requires the estimation of d, the dimension of the central space, since we know that the true dimension is 1, we used 1 in the selection procedure. Thus the problem with COP does not stem from model misspecification and we are surprised by its poor performance. Table 5.1 shows that COP tends to correctly select predictors 1 and 2 but miss predictors 3 and 4.

MFVS-S and MVFS-D perform quite well in the linear setting. Regardless of transformation used we see false negative rates of nearly zero in most cases. For both methods the best performance, in terms of lowest FPR is achieved when using no transformation or when using the normalizing transformation. The worst performance happens when the response transformation is used, and this is especially prominent when p is large and when ρ is large. Since when the response transformation is used the variable selection procedure run twice, it is not surprising to see a higher rate of false positives. We posit that this higher

rate of false positives is justified provided that the response transformation enhances selection when a heteroscedastic error is present. However, future research toward lowering the false positive rate associated with the response transformation may be beneficial.

Finally, in model 1 we also notice that there is some relationship between false negative rate, correlation and predictor size. As predictor correlation increases, the false negative rate does as well. This effect is more pronounced when p is larger, for example in Figure 5.5. The effect is also more pronounced for the response transformation. Unfortunately, our methods are not the only ones to have this problem as multicollinearity plagues many variable selection and modeling techniques.

Method	f_1	f_2	f_3	f_4	FPR	FNR	$\lvert\mathcal{F}\rvert$
MFVS-D (All)	0.95	1.00	1.00	0.87	0.06	0.04	6.35
MFVS-D (None)	0.99	1.00	1.00	0.95	0.01	0.01	4.35
MFVS-D (Norm)	0.98	1.00	1.00	0.97	0.01	0.01	4.26
MFVS-D (Symm Link)	0.99	1.00	1.00	0.96	0.02	0.01	4.89
MFVS-D (Response)	0.98	1.00	1.00	0.88	0.05	0.04	6.21
MFVS-S (All)	0.93	1.00	1.00	0.87	0.09	0.05	8.01
MFVS-S (None)	1.00	1.00	1.00	0.92	0.00	0.02	3.92
MFVS-S (Norm)	1.00	1.00	1.00	0.93	0.00	0.02	3.93
MFVS-S (Symm Link)	1.00	1.00	1.00	0.91	0.00	0.02	3.97
MFVS-S (Response)	0.97	1.00	1.00	0.91	0.08	0.03	7.68
Lasso	1.00	1.00	1.00	1.00	0.03	0.00	5.42
COP	0.99	1.00	0.13	0.23	0.15	0.41	9.10

Table 5.1: Linear model: simulation results for $p = 50$ and $\rho = 0.3$ for each method (transformations in parentheses).

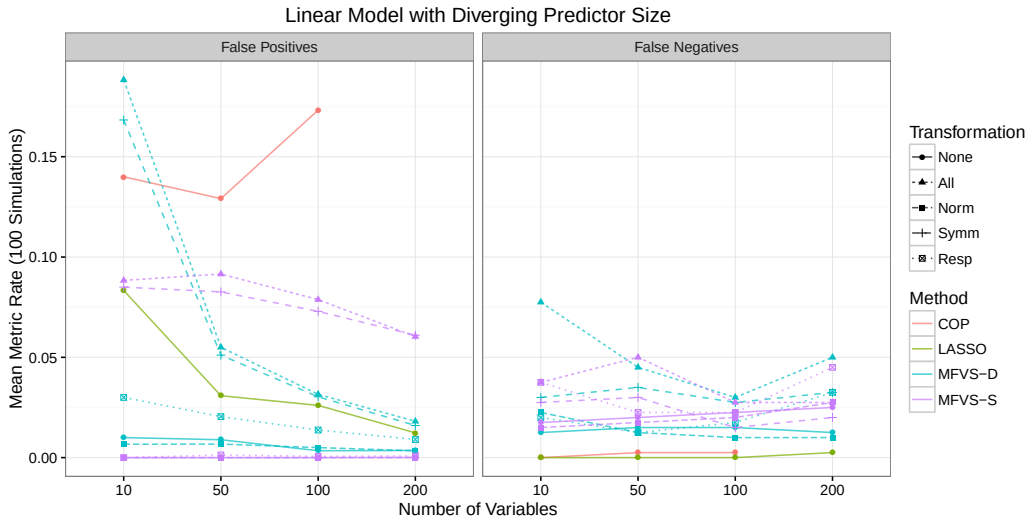

Figure 5.1: Linear model with diverging predictor size: mean false positive and false negative rates over 100 simulations.

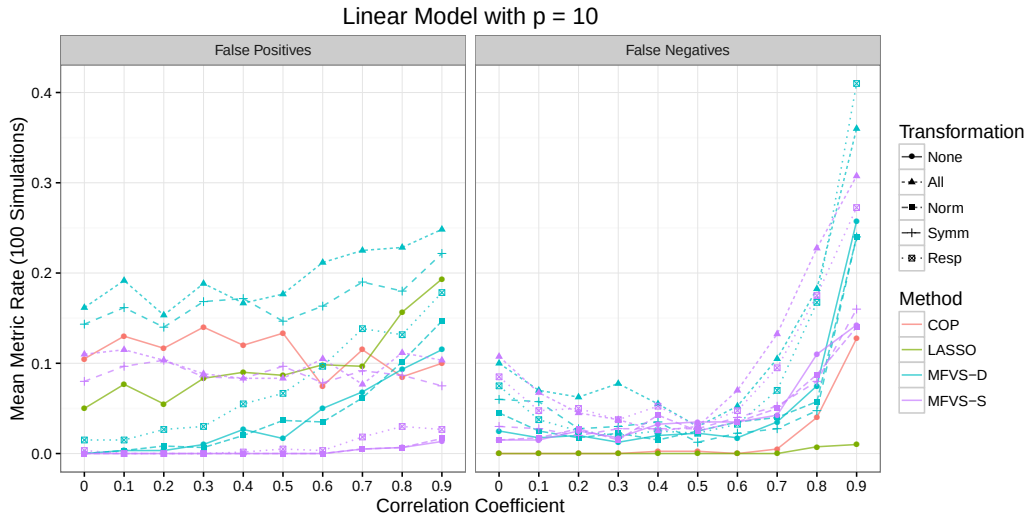

Figure 5.2: Linear model with varying correlation and $p = 10$: mean false positive and false negative rates over 100 simulations.

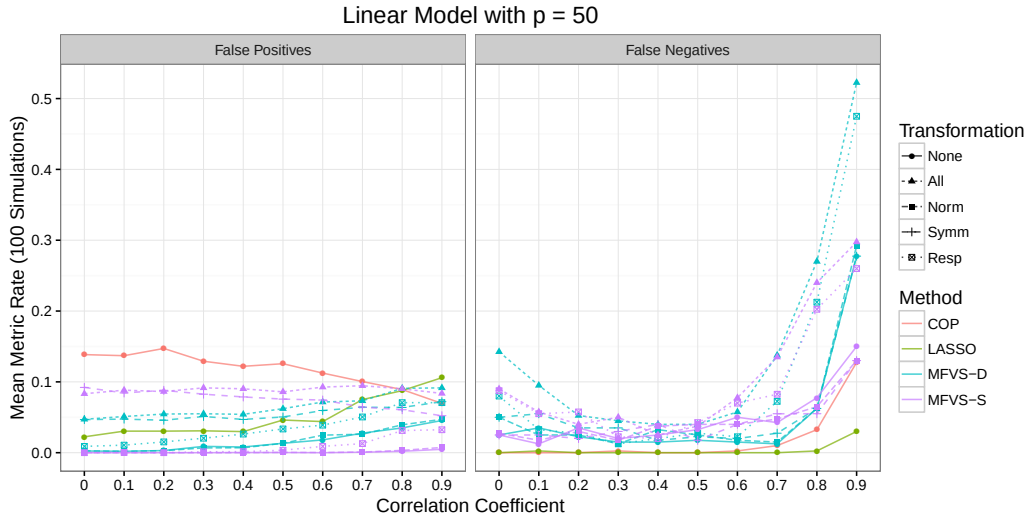

Figure 5.3: Linear model with varying correlation and $p = 50$: mean false positive and false negative rates over 100 simulations.

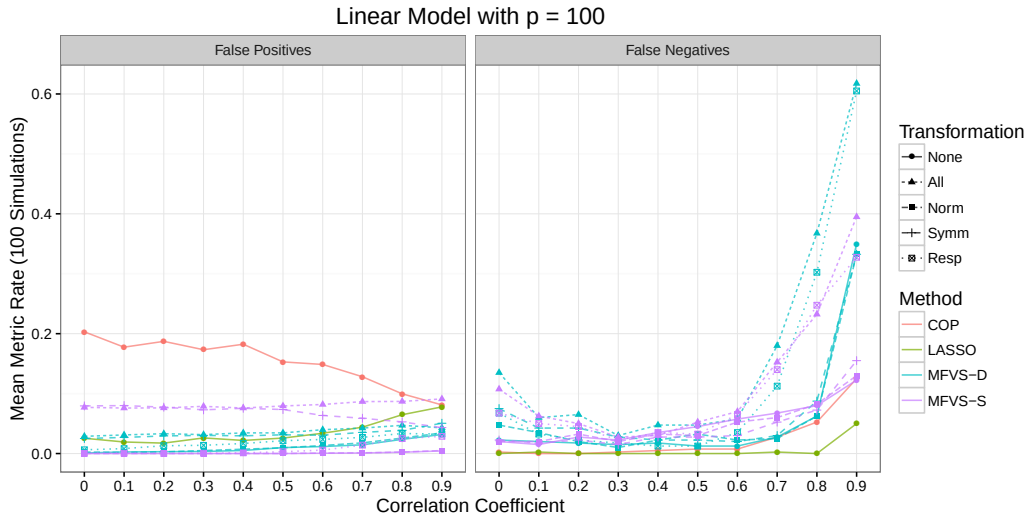

Figure 5.4: Linear model with varying correlation and $p = 100$: mean false positive and false negative rates over 100 simulations.

Method	f_1	f_2	f_3	f_4	FPR	FNR	$\lvert\mathcal{F}\rvert$
MFVS-D (All)	0.96	1.00	1.00	0.84	0.02	0.05	7.39
MFVS-D (None)	0.98	1.00	1.00	0.97	0.00	0.01	4.65
MFVS-D (Norm)	0.98	1.00	1.00	0.98	0.00	0.01	4.63
MFVS-D (Symm Link)	0.99	1.00	1.00	0.88	0.01	0.03	5.64
MFVS-D (Response)	0.98	1.00	1.00	0.89	0.02	0.03	7.01
MFVS-S (All)	1.00	1.00	1.00	0.89	0.06	0.03	15.72
MFVS-S (None)	1.00	1.00	1.00	0.90	0.00	0.03	3.92
MFVS-S (Norm)	1.00	1.00	1.00	0.89	0.00	0.03	3.90
MFVS-S (Symm Link)	1.00	1.00	1.00	0.82	0.00	0.04	3.99
MFVS-S (Response)	1.00	1.00	1.00	0.92	0.06	0.02	15.88
Lasso	1.00	1.00	1.00	0.99	0.01	0.00	6.41
COP	0.00	0.00	0.00	0.00	NA	NA	1.00

Table 5.2: Linear model: simulation results for $p = 200$ and $\rho = 0.3$ for each method (transformations in parentheses).

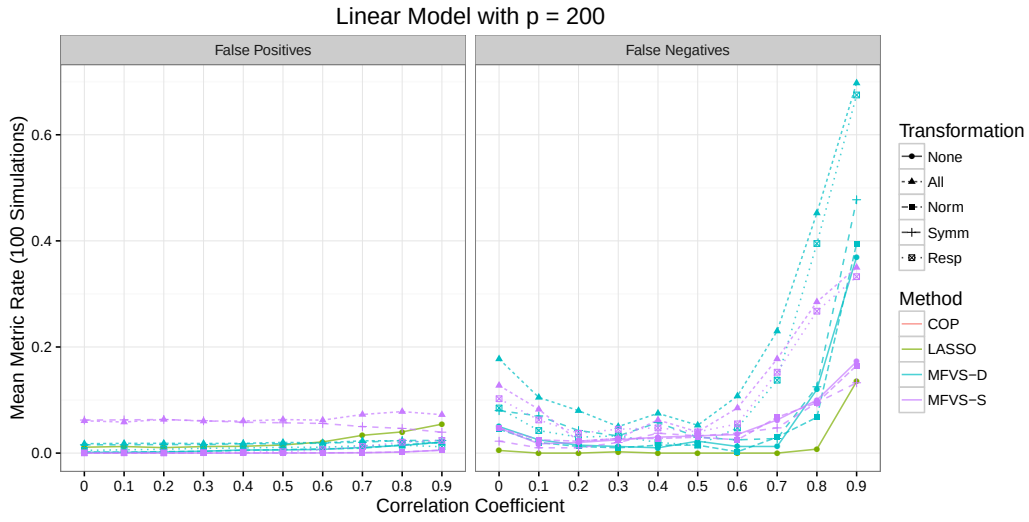

Figure 5.5: Linear model with varying correlation and $p = 200$: mean false positive and false negative rates over 100 simulations.

5.1.2 Model 2: Nonlinear

The results for model 2 are shown in Tables 5.3 and 5.4 and in Figures 5.6, 5.7, 5.8, 5.9 and 5.10. Recall that model 2 is:

$$Y = X_2^3 + \exp(X_2) + (X_3 + X_4)^{1.5} + \varepsilon$$

In model 2, the nonlinear model, we see results similar to the results from the linear model. This indicates that our methods work as well for nonlinear models as for linear, thus confirming our theoretical ideas.

In this model we first begin to see a difference between the stepwise procedure and the penalized procedure. As the predictor variable correlation increases, the false negative rate of the stepwise procedure increases as well. This indicates that the stepwise procedure may not be the best choice when correlation is above $\rho = 0.4$, and the problem is amplified as p increases. The penalized method does not appear to suffer from this problem. In particular, the stepwise procedure misses X_2 in 2% to 11% of the simulations.

LASSO also misses X_2, up to 5% of the time. This is not a high number of misses, but the same miss rate for MFVS-D with no transformation is 0% and with all transformations it is 2%. Since LASSO and the Dantzig selector are so similar, we have some practical evidence that while the linear L1 penalization method works well in the nonlinear setting, using the nonlinear (GAM) model for cross validation may increase accuracy.

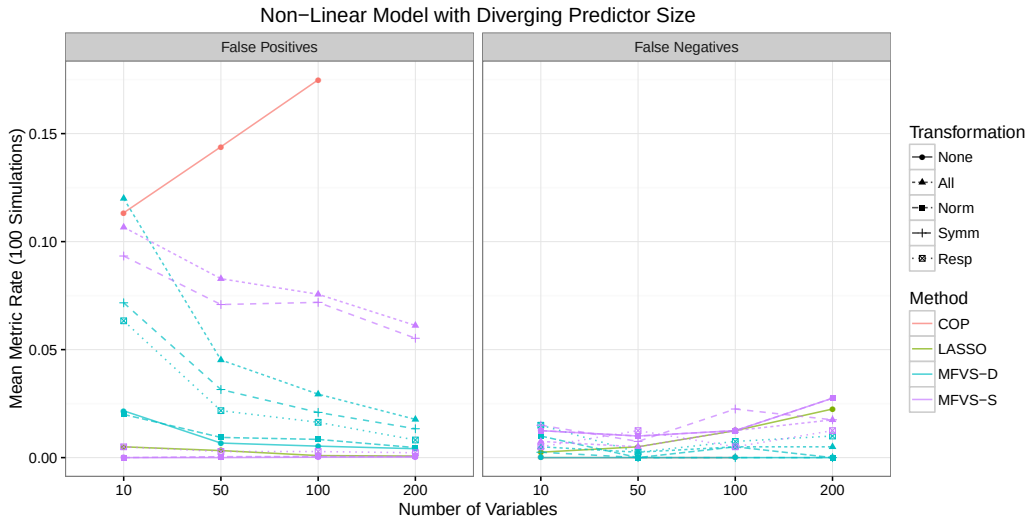

Figure 5.6: Non-Linear model with diverging predictor size: mean false positive and false negative rates over 100 simulations.

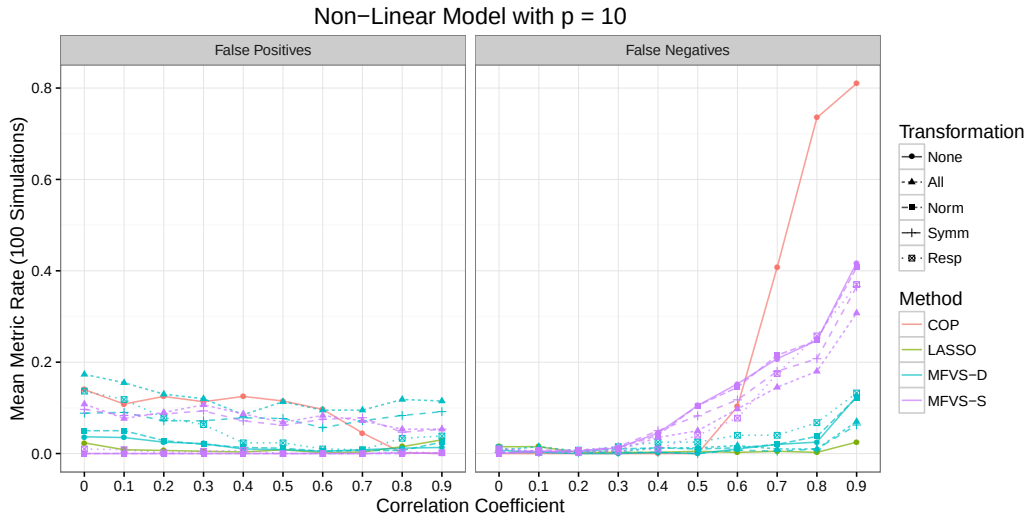

Figure 5.7: Non-Linear model with varying correlation and $p = 10$: mean false positive and false negative rates over 100 simulations.

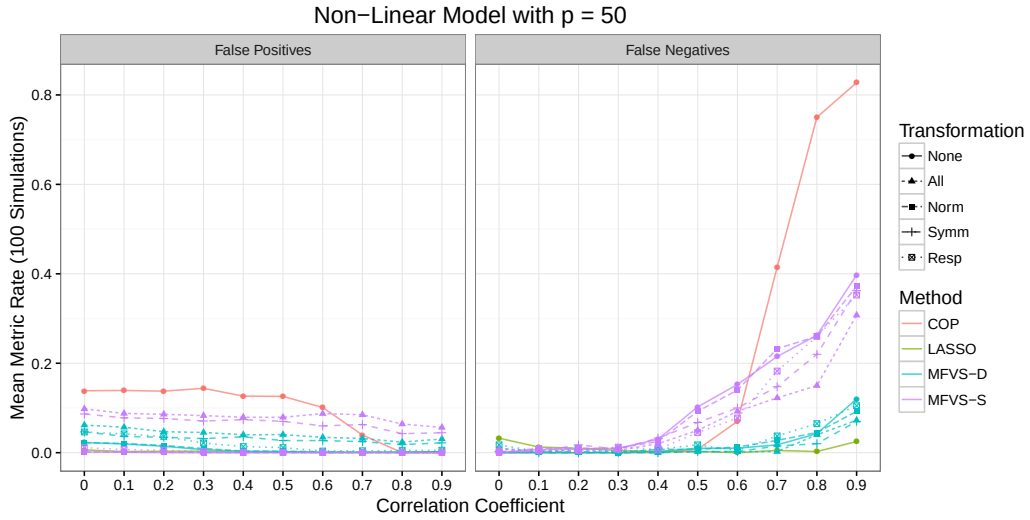

Figure 5.8: Non-Linear model with varying correlation and $p = 50$: mean false positive and false negative rates over 100 simulations.

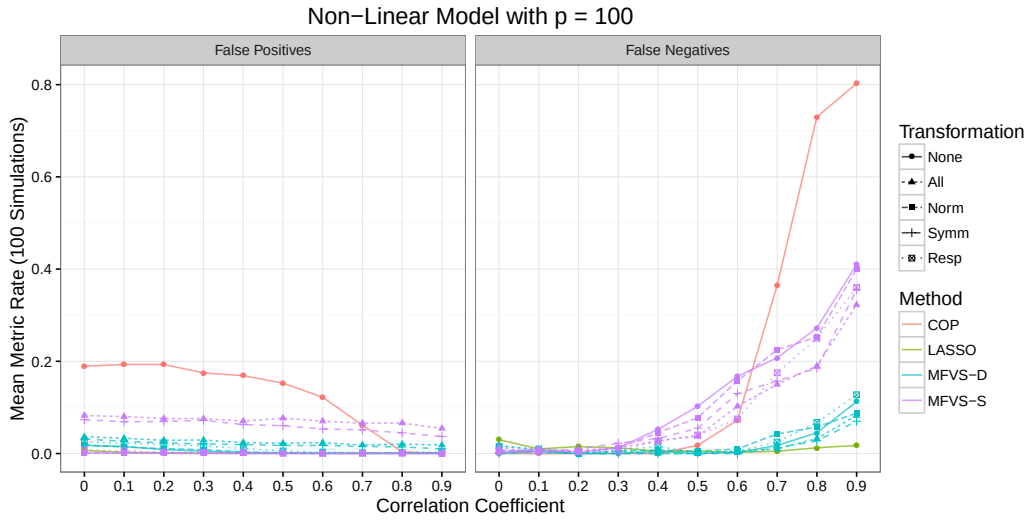

Figure 5.9: Non-Linear model with varying correlation and $p = 100$: mean false positive and false negative rates over 100 simulations.

Method	f_1	f_2	f_3	f_4	FPR	FNR	$\lvert\mathcal{F}\rvert$
MFVS-D (All)	1.00	0.99	1.00	1.00	0.05	0.00	6.23
MFVS-D (None)	1.00	1.00	1.00	1.00	0.01	0.00	4.31
MFVS-D (Norm)	1.00	1.00	1.00	1.00	0.01	0.00	4.43
MFVS-D (Symm Link)	1.00	1.00	1.00	0.99	0.02	0.00	4.99
MFVS-D (Response)	1.00	1.00	1.00	1.00	0.03	0.00	5.49
MFVS-S (All)	1.00	0.98	1.00	1.00	0.08	0.01	7.79
MFVS-S (None)	1.00	0.96	1.00	1.00	0.00	0.01	3.96
MFVS-S (Norm)	1.00	0.96	1.00	1.00	0.00	0.01	3.98
MFVS-S (Symm Link)	1.00	0.95	1.00	1.00	0.00	0.01	4.08
MFVS-S (Response)	1.00	0.97	1.00	1.00	0.07	0.01	7.23
Lasso	1.00	0.98	1.00	1.00	0.00	0.01	4.13
COP	0.99	1.00	0.13	0.23	0.15	0.41	9.10

Table 5.3: Non-Linear model: simulation results for $p = 50$ and $\rho = 0.3$ for each method (transformations in parentheses).

Method	f_1	f_2	f_3	f_4	FPR	FNR	$\lvert\mathcal{F}\rvert$
MFVS-D (All)	1.00	0.98	1.00	1.00	0.02	0.01	7.49
MFVS-D (None)	1.00	1.00	1.00	1.00	0.00	0.00	4.81
MFVS-D (Norm)	1.00	1.00	1.00	1.00	0.00	0.00	4.91
MFVS-D (Symm Link)	1.00	0.98	0.99	0.99	0.01	0.01	5.56
MFVS-D (Response)	1.00	1.00	1.00	1.00	0.01	0.00	6.70
MFVS-S (All)	1.00	0.93	1.00	1.00	0.06	0.02	15.92
MFVS-S (None)	1.00	0.89	1.00	1.00	0.00	0.03	3.94
MFVS-S (Norm)	1.00	0.89	1.00	1.00	0.00	0.03	4.00
MFVS-S (Symm Link)	1.00	0.95	1.00	1.00	0.00	0.01	4.39
MFVS-S (Response)	1.00	0.93	1.00	1.00	0.06	0.02	14.76
Lasso	0.99	0.95	1.00	0.97	0.00	0.02	4.05
COP	0.00	0.00	0.00	0.00	NA	NA	1.00

Table 5.4: Non-Linear model: simulation results for $p = 200$ and $\rho = 0.3$ for each method (transformations in parentheses).

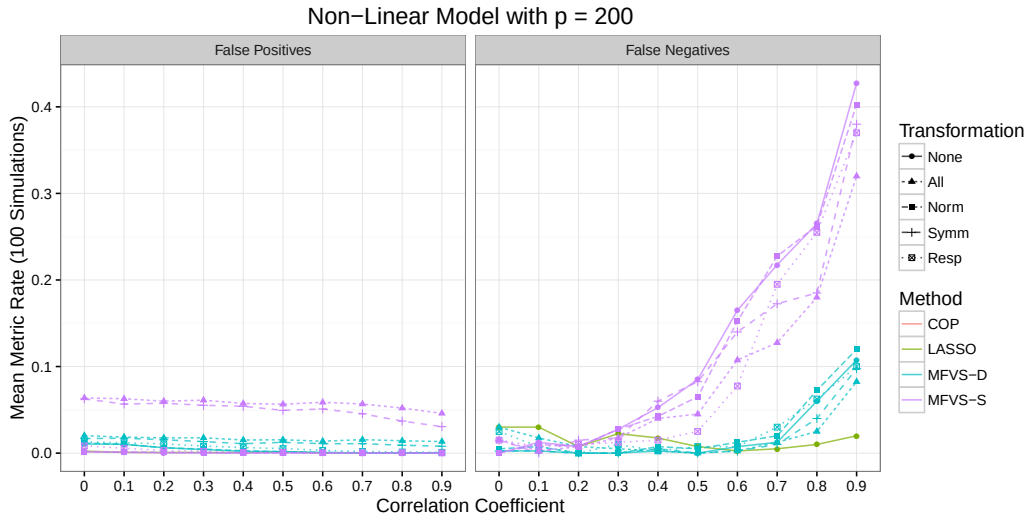

Figure 5.10: Non-Linear model with varying correlation and $p = 200$: mean false positive and false negative rates over 100 simulations.

5.1.3 Model 3: Nonlinear and Non-Normal

The results for model 3 are shown in Tables 5.5 and 5.6 and in Figures 5.11, 5.12, 5.13, 5.14 and 5.15. Recall that model 3 is:

$$Y = X_2^3 + \exp(X_2) + (X_3 + X_4)^{1.5} + \varepsilon$$

For model 3, the non-linear and non-normal model MFVS-D is much more accurate compared to LASSO, which is very inaccurate with False Negative rates above 50%. This shows that violating the assumption of normally distributed predictors can have very negative consequences. We also see the same trend as in the previous models that the penalized procedure generally performs better than the stepwise procedure and that the stepwise procedure tends to have higher false negative rates as the correlation increases.

There are two particularly interesting results in this set of simulations. First, the normalizing transformation does not lead to the lowest false negative rates across the board; however there is no one transformation that is the best in all cases under model 3. Especially when correlation is high, using all three transformations tends to lead to the lowest false negative rate. The second interesting results is that using no transformation with MFVS-D still leads to accurate results, especially when compared to LASSO. This indicates, again, that using a linear model criterion in cross validation leads to decreased accuracy.

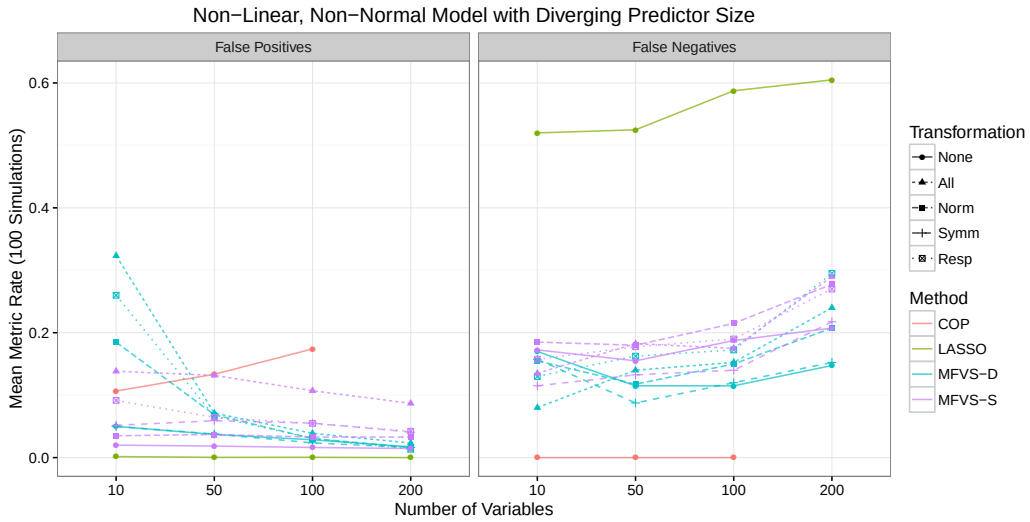

Figure 5.11: Non-Linear, Non-Normal model with diverging predictor size: mean false positive and false negative rates over 100 simulations.

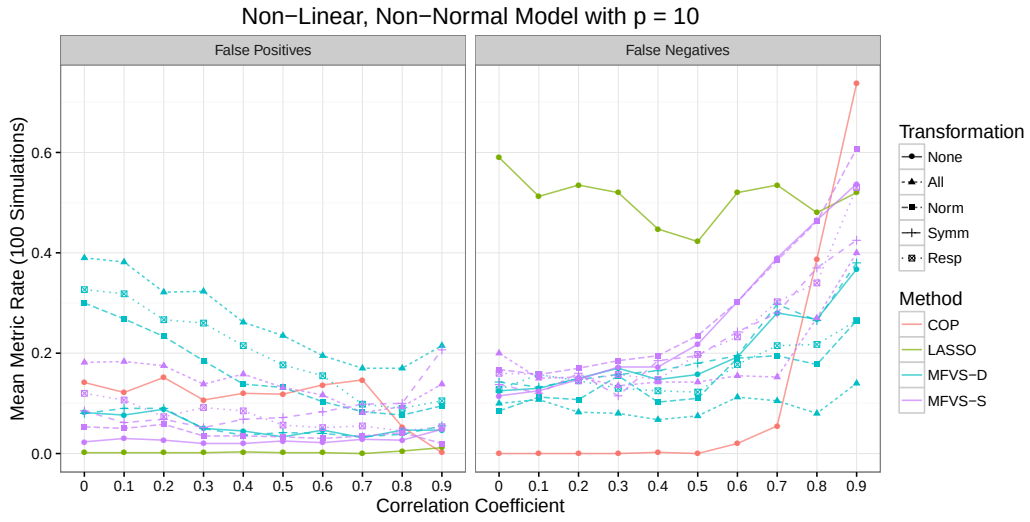

Figure 5.12: Non-Linear, Non-Normal model with varying correlation and $p = 10$: mean false positive and false negative rates over 100 simulations.

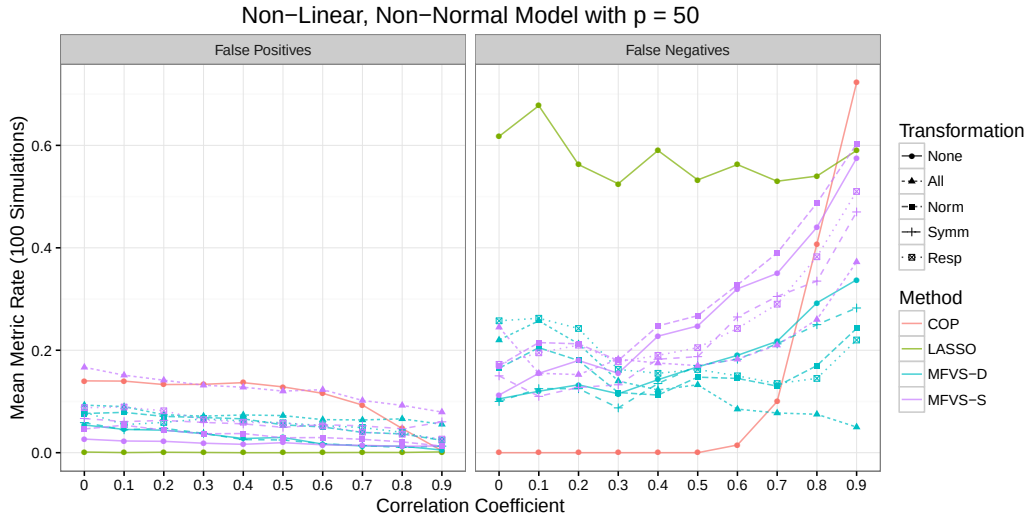

Figure 5.13: Non-Linear, Non-Normal model with varying correlation and $p = 50$: mean false positive and false negative rates over 100 simulations.

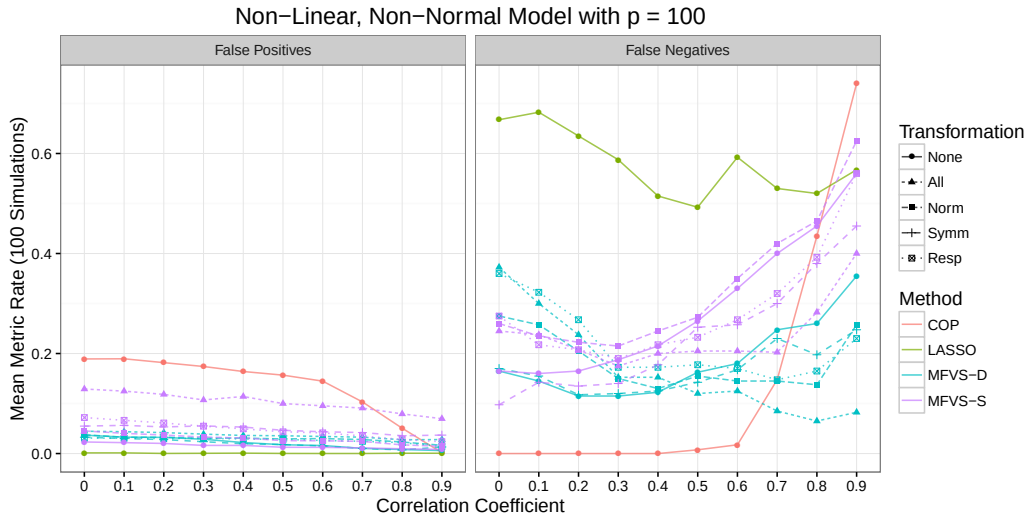

Figure 5.14: Non-Linear, Non-Normal model with varying correlation and $p = 100$: mean false positive and false negative rates over 100 simulations.

Method	f_1	f_2	f_3	f_4	FPR	FNR	$\lvert\mathcal{F}\rvert$
MFVS-D (All)	0.96	0.97	0.83	0.68	0.07	0.14	6.76
MFVS-D (None)	0.95	0.97	0.82	0.80	0.04	0.12	5.25
MFVS-D (Norm)	0.97	0.98	0.81	0.77	0.07	0.12	6.69
MFVS-D (Symm Link)	0.93	0.97	0.80	0.65	0.07	0.16	6.39
MFVS-D (Response)	0.97	0.99	0.86	0.83	0.04	0.09	5.39
MFVS-S (All)	0.92	0.97	0.77	0.61	0.13	0.18	9.33
MFVS-S (None)	0.94	0.96	0.69	0.79	0.02	0.15	4.23
MFVS-S (Norm)	0.95	0.94	0.69	0.70	0.04	0.18	4.98
MFVS-S (Symm Link)	0.92	0.96	0.74	0.67	0.06	0.18	6.25
MFVS-S (Response)	0.94	0.97	0.77	0.79	0.06	0.13	6.19
Lasso	0.57	0.61	0.40	0.32	0.00	0.53	1.92
COP	0.99	1.00	0.13	0.23	0.15	0.41	9.10

Table 5.5: Non-Linear, Non-Normal model: simulation results for $p = 50$ and $\rho = 0.3$ for each method (transformations in parentheses).

Method	f_1	f_2	f_3	f_4	FPR	FNR	$\lvert\mathcal{F}\rvert$
MFVS-D (All)	0.91	0.90	0.65	0.58	0.02	0.24	7.62
MFVS-D (None)	0.94	0.93	0.77	0.77	0.02	0.15	6.79
MFVS-D (Norm)	0.90	0.93	0.76	0.58	0.02	0.21	6.31
MFVS-D (Symm Link)	0.87	0.88	0.62	0.45	0.01	0.29	5.45
MFVS-D (Response)	0.95	0.93	0.76	0.75	0.02	0.15	6.57
MFVS-S (All)	0.83	0.86	0.59	0.56	0.09	0.29	19.88
MFVS-S (None)	0.92	0.89	0.66	0.70	0.01	0.21	6.03
MFVS-S (Norm)	0.87	0.87	0.62	0.53	0.03	0.28	9.34
MFVS-S (Symm Link)	0.86	0.89	0.67	0.50	0.04	0.27	11.20
MFVS-S (Response)	0.92	0.91	0.60	0.70	0.04	0.22	11.15
Lasso	0.50	0.47	0.38	0.23	0.00	0.60	1.59
COP	0.00	0.00	0.00	0.00	NA	NA	1.00

Table 5.6: Non-Linear, Non-Normal model: simulation results for $p = 200$ and $\rho = 0.3$ for each method (transformations in parentheses).

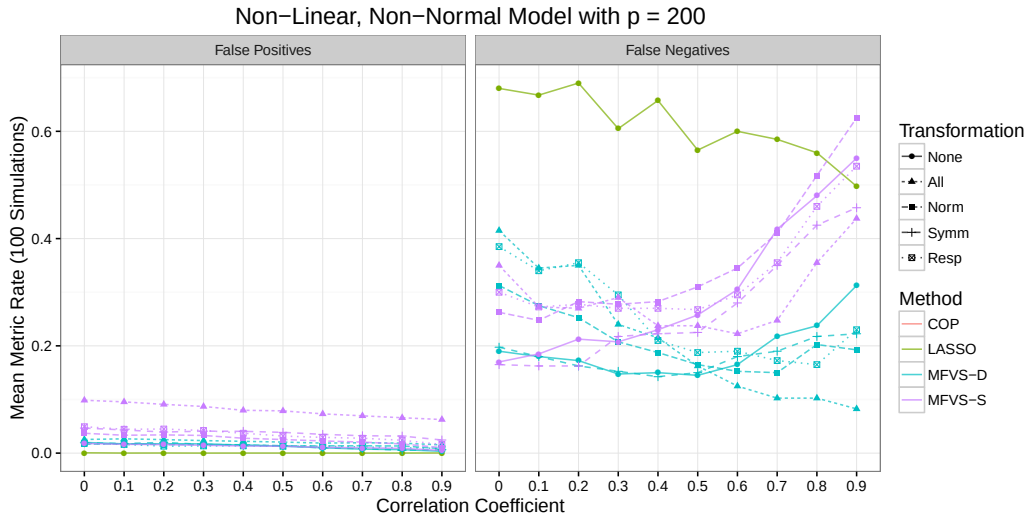

Figure 5.15: Non-Linear, Non-Normal model with varying correlation and $p = 200$: mean false positive and false negative rates over 100 simulations.

5.1.4 Model 4: heteroscedastic Error or Central Space

The results for model 4 are shown in Tables 5.7 and 5.8 and in Figures 5.16, 5.17, 5.18, 5.19 and 5.20. Recall that model 4 is:

$$Y = X_1 + X_2 + X_3 + X_4\varepsilon$$

We saw in the previous 3 cases and we will see in the next case that the response transformation can help in several cases in addition to what it was designed for, which is to select variables in the variance component, i.e. heteroscedastic error. In this case the response transformation also helps, but it unfortunately does not ensure that the variable in the variance component, X_4 is always selected. When using the response transformation X_4 is only selected by MFVS-S and MFVS-D about 20 to 30% of the time. This is, fortunately, much better than the methods without the response transformation which select X_4 nearly 0% of the time. By comparison, COP, which was designed to be able to select variables in the variance component, was only able to select X_4 about 17% of the time. This indicates that variable selection in the presence of heteroscedastic error can be quite challenging, yet we are able to improve accuracy with the response transformation.

Since model 3 is essentially a linear model with one variable in the variance component, the selection results for X_1, X_2 and X_3 are about the same as the results for model 1, the linear model.

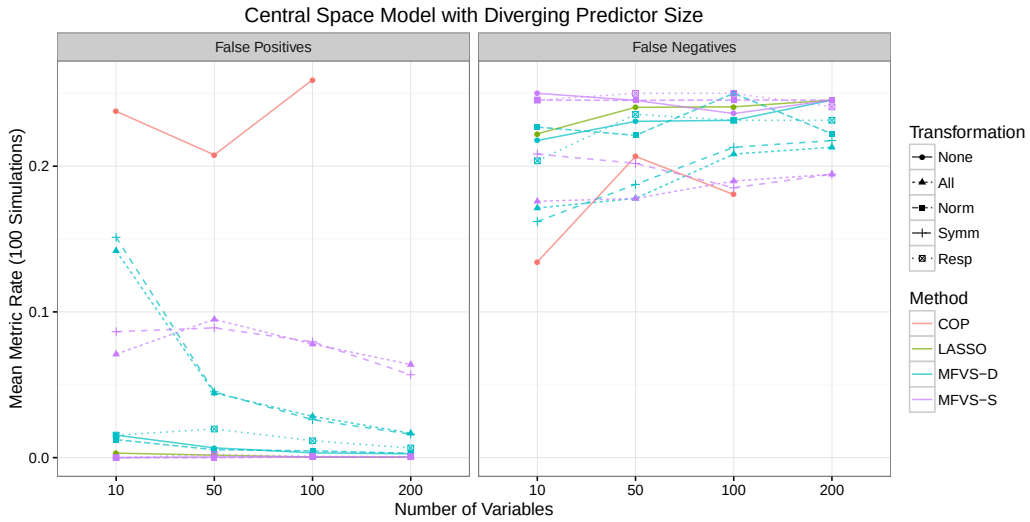

Figure 5.16: Central Space model with diverging predictor size: mean false positive and false negative rates over 100 simulations.

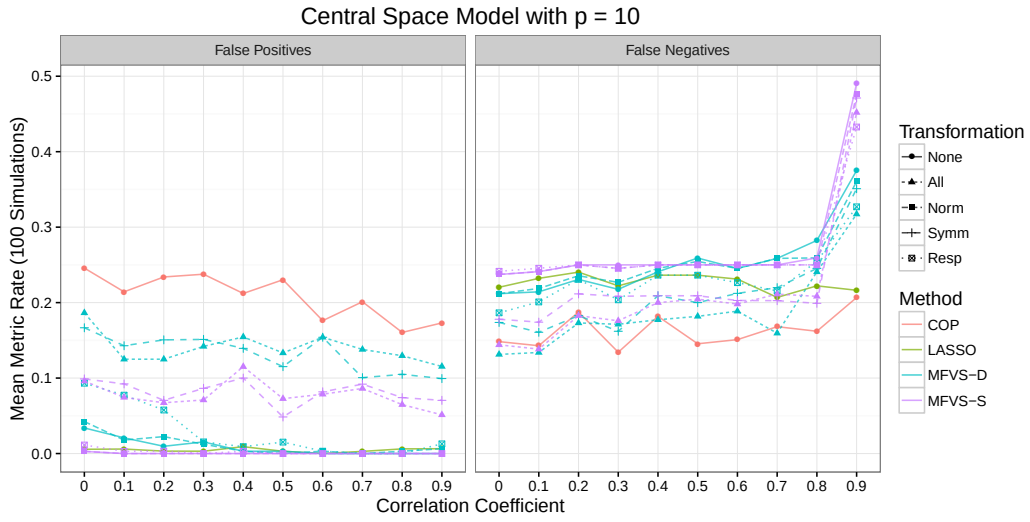

Figure 5.17: Central Space model with varying correlation and $p = 10$: mean false positive and false negative rates over 100 simulations.

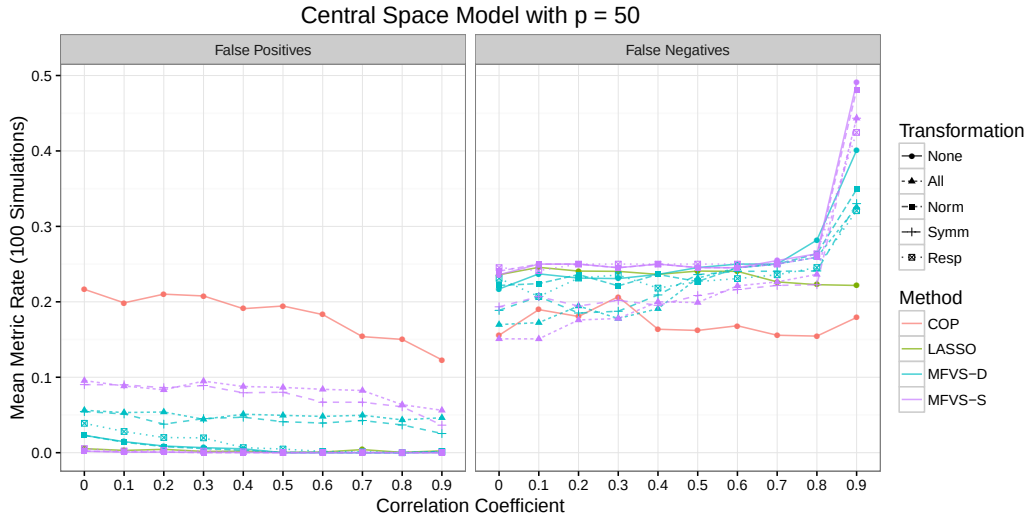

Figure 5.18: Central Space model with varying correlation and $p = 50$: mean false positive and false negative rates over 100 simulations.

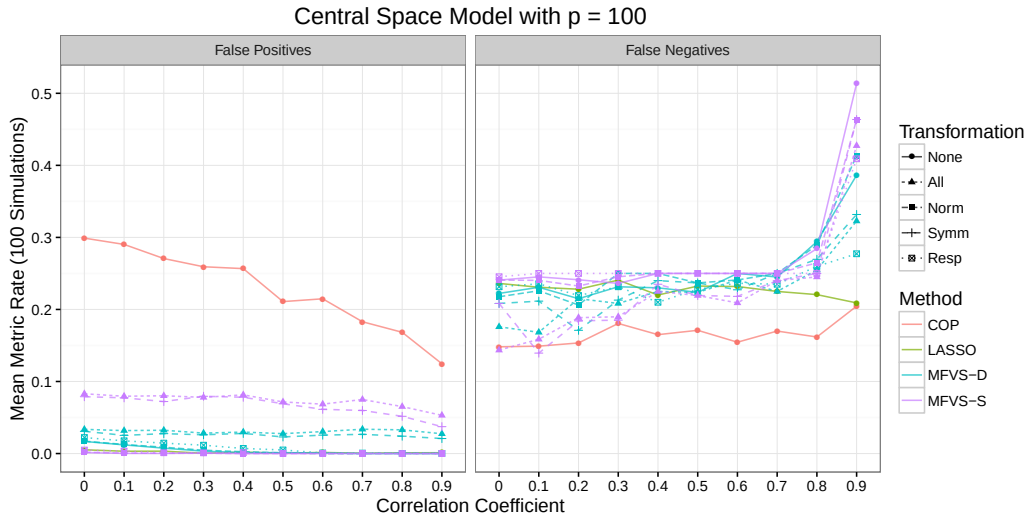

Figure 5.19: Central Space model with varying correlation and $p = 100$: mean false positive and false negative rates over 100 simulations.

Method	f_1	f_2	f_3	f_4	FPR	FNR	$\lvert\mathcal{F}\rvert$
MFVS-D (All)	1.00	1.00	1.00	0.28	0.04	0.18	5.28
MFVS-D (None)	1.00	1.00	1.00	0.09	0.01	0.23	3.40
MFVS-D (Norm)	1.00	1.00	1.00	0.13	0.01	0.22	3.40
MFVS-D (Symm Link)	1.00	1.00	1.00	0.06	0.02	0.23	3.98
MFVS-D (Response)	1.00	1.00	1.00	0.26	0.04	0.19	5.26
MFVS-S (All)	1.00	1.00	1.00	0.28	0.09	0.18	7.57
MFVS-S (None)	1.00	1.00	1.00	0.02	0.00	0.24	3.04
MFVS-S (Norm)	1.00	1.00	1.00	0.02	0.00	0.24	3.02
MFVS-S (Symm Link)	1.00	1.00	1.00	0.00	0.00	0.25	3.11
MFVS-S (Response)	1.00	1.00	1.00	0.21	0.09	0.20	7.13
Lasso	1.00	1.00	1.00	0.04	0.00	0.24	3.11
COP	1.00	1.00	1.00	0.17	0.20	0.21	12.51

Table 5.7: Central Space model: simulation results for $p = 50$ and $\rho = 0.3$ for each method (transformations in parentheses).

Method	f_1	f_2	f_3	f_4	FPR	FNR	$\lvert\mathcal{F}\rvert$
MFVS-D (All)	1.00	1.00	1.00	0.16	0.02	0.21	6.30
MFVS-D (None)	1.00	1.00	0.98	0.05	0.00	0.24	3.55
MFVS-D (Norm)	1.00	1.00	1.00	0.11	0.00	0.22	3.68
MFVS-D (Symm Link)	1.00	1.00	1.00	0.07	0.01	0.23	4.45
MFVS-D (Response)	1.00	1.00	1.00	0.11	0.01	0.22	6.05
MFVS-S (All)	1.00	1.00	1.00	0.20	0.06	0.20	15.68
MFVS-S (None)	1.00	1.00	1.00	0.02	0.00	0.24	3.09
MFVS-S (Norm)	1.00	1.00	1.00	0.02	0.00	0.24	3.07
MFVS-S (Symm Link)	1.00	1.00	1.00	0.02	0.00	0.24	3.16
MFVS-S (Response)	1.00	1.00	1.00	0.20	0.06	0.20	14.59
Lasso	1.00	1.00	1.00	0.02	0.00	0.24	3.14
COP	0.00	0.00	0.00	0.00	NA	NA	1.00

Table 5.8: Central Space model: simulation results for $p = 200$ and $\rho = 0.3$ for each method (transformations in parentheses).

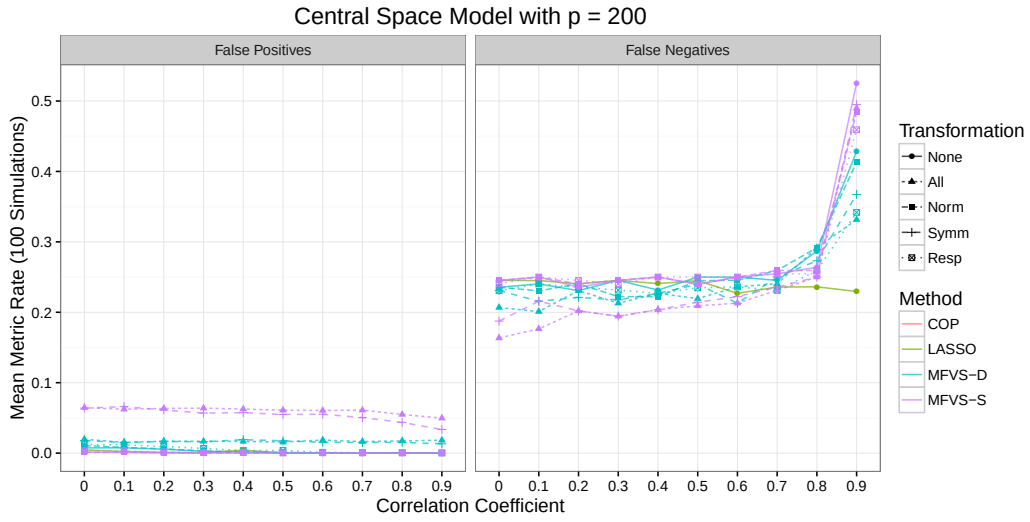

Figure 5.20: Central Space model with varying correlation and $p = 200$: mean false positive and false negative rates over 100 simulations.

5.1.5 Model 5: Symmetric Link Function

The results for model 5 are shown in Tables 5.9 and 5.10 and in Figures 5.21, 5.22, 5.23, 5.24 and 5.25. Recall that model 5 is:

$$Y = X_1 + 0.75X_2^2 + X_3 + 0.5X_4^4 + \varepsilon$$

For model 5, two of the four predictor variables (X_2 and X_4) had a symmetric relationship with the response and it should be no surprise that the transformation for symmetric link functions is a great help to both the stepwise method and the penalized method in the case of symmetric link function. In this case, using all three transformations or using only the symmetry transformation lead to increased accuracy compared to using no transformation or using only the normalizing transformation.

Interestingly, again, the response transformation leads to more accurate results than using no transformation, even though the response transformation was not designed to correct for the problem of a symmetric link function. However, again, the response transformation does results in a higher number of false positives.

We also see that LASSO is unable to select any predictors, even though X_1 and X_3 do not have a symmetric relationship with the response. The false negative rates for LASSO are frequently 100%. Again, we do not see this to be the case with MFVS-D with no predictors, indicating again that the model-free

cross validation is crucial to model-free variable selection. It is clear from these simulations that using a linear model can have very dire consequences if there is a symmetric relationship between the response and any of the predictors.

Method	f_1	f_2	f_3	f_4	FPR	FNR	$\lvert\mathcal{F}\rvert$
MFVS-D (All)	0.71	0.72	0.70	1.00	0.06	0.22	5.74
MFVS-D (None)	0.78	0.46	0.80	0.88	0.09	0.27	7.27
MFVS-D (Norm)	0.78	0.32	0.71	0.83	0.08	0.34	6.45
MFVS-D (Symm Link)	0.74	0.77	0.74	1.00	0.05	0.19	5.66
MFVS-D (Response)	0.79	0.44	0.74	0.92	0.09	0.28	6.91
MFVS-S (All)	0.71	0.70	0.77	1.00	0.13	0.20	8.97
MFVS-S (None)	0.74	0.16	0.70	0.48	0.06	0.48	4.97
MFVS-S (Norm)	0.75	0.12	0.70	0.37	0.06	0.52	4.84
MFVS-S (Symm Link)	0.75	0.79	0.70	1.00	0.04	0.19	5.28
MFVS-S (Response)	0.82	0.19	0.70	0.76	0.11	0.38	7.34
Lasso	0.08	0.03	0.08	0.00	0.00	0.95	0.19
COP	0.99	1.00	0.13	0.23	0.15	0.41	9.10

Table 5.9: Symmetric Link model: simulation results for $p = 50$ and $\rho = 0.3$ for each method (transformations in parentheses).

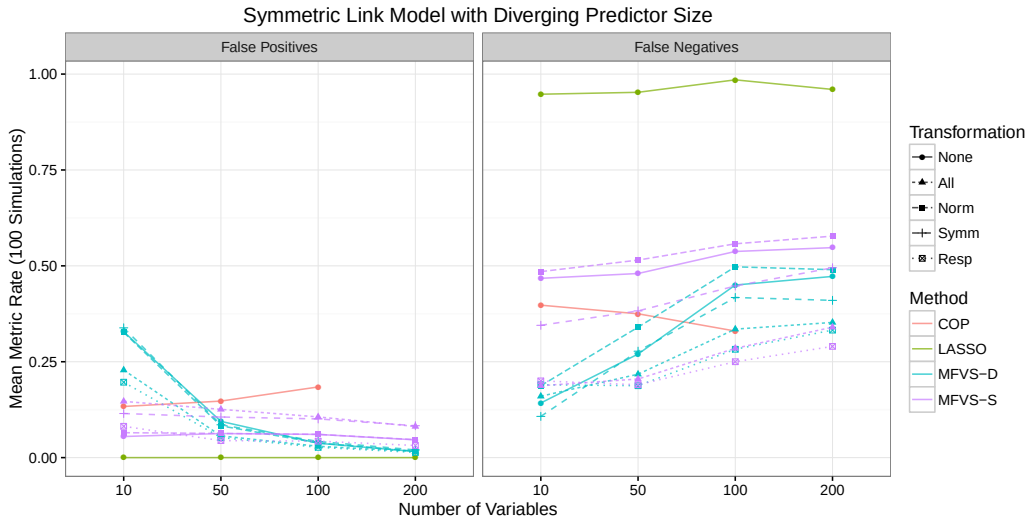

Figure 5.21: Symmetric Link model with diverging predictor size: mean false positive and false negative rates over 100 simulations.

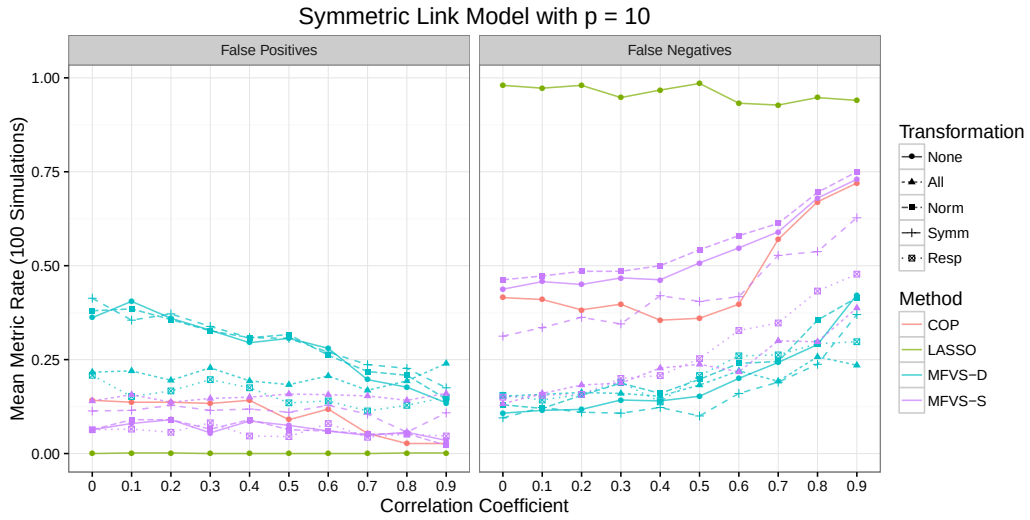

Figure 5.22: Symmetric Link model with varying correlation and $p = 10$: mean false positive and false negative rates over 100 simulations.

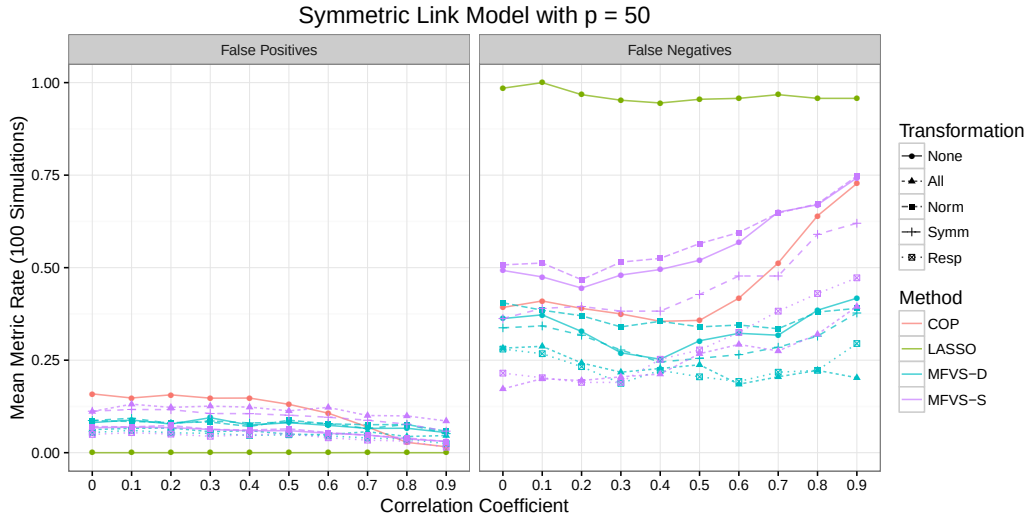

Figure 5.23: Symmetric Link model with varying correlation and $p = 50$: mean false positive and false negative rates over 100 simulations.

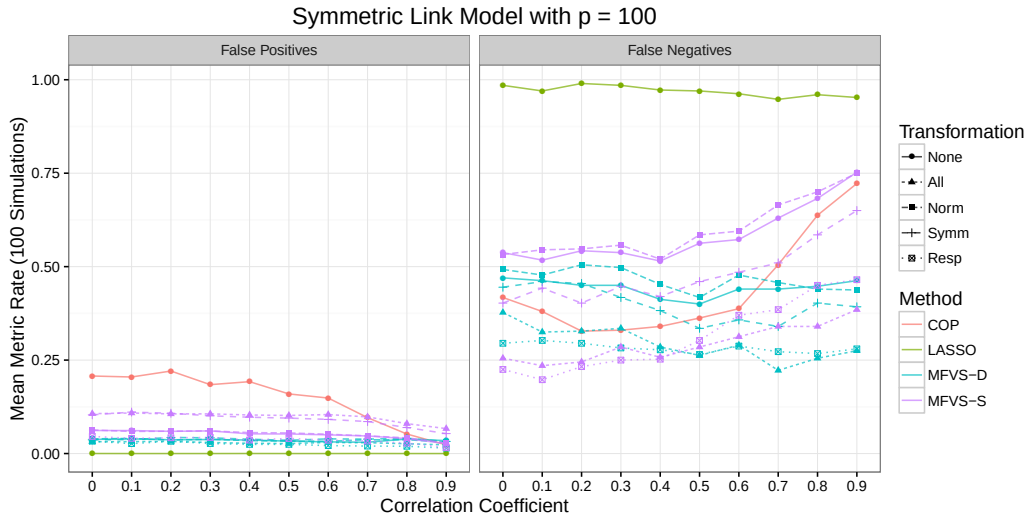

Figure 5.24: Symmetric Link model with varying correlation and $p = 100$: mean false positive and false negative rates over 100 simulations.

Method	f_1	f_2	f_3	f_4	FPR	FNR	$\lvert\mathcal{F}\rvert$
MFVS-D (All)	0.52	0.53	0.54	1.00	0.02	0.35	5.79
MFVS-D (None)	0.64	0.16	0.61	0.70	0.02	0.47	5.23
MFVS-D (Norm)	0.58	0.18	0.65	0.63	0.02	0.49	5.29
MFVS-D (Symm Link)	0.59	0.54	0.54	1.00	0.01	0.33	5.32
MFVS-D (Response)	0.60	0.21	0.67	0.88	0.02	0.41	6.33
MFVS-S (All)	0.61	0.53	0.50	1.00	0.08	0.34	18.56
MFVS-S (None)	0.63	0.08	0.72	0.38	0.05	0.55	10.88
MFVS-S (Norm)	0.61	0.07	0.72	0.29	0.05	0.58	10.88
MFVS-S (Symm Link)	0.67	0.55	0.62	1.00	0.03	0.29	9.04
MFVS-S (Response)	0.59	0.11	0.70	0.62	0.08	0.49	18.32
Lasso	0.06	0.01	0.07	0.02	0.00	0.96	0.16
COP	0.00	0.00	0.00	0.00	NA	NA	1.00

Table 5.10: Symmetric Link model: simulation results for $p = 200$ and $\rho = 0.3$ for each method (transformations in parentheses).

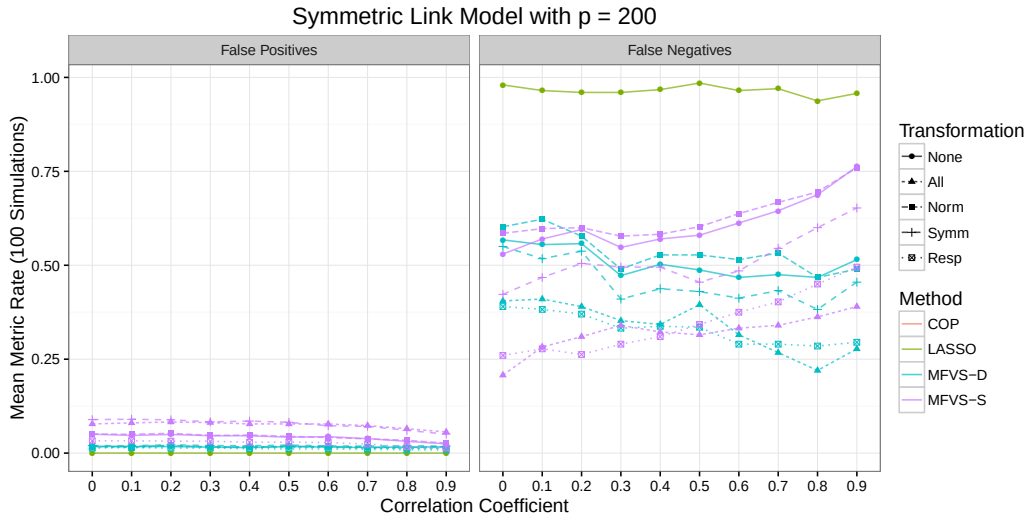

Figure 5.25: Symmetric Link model with varying correlation and $p = 200$: mean false positive and false negative rates over 100 simulations.

5.1.6 Model 6: Combination

The results for model 6 are shown in Tables 5.11 and 5.12 and in Figures 5.26, 5.27, 5.28, 5.29 and 5.30. Recall that model 6 is:

$$Y = X_1^2 + X_2 + X_3^4 + X_4\varepsilon$$

The final model is designed to test MFVS-S and MFVS-D when there is more than one factor contributing to the nonlinearity of the relationship between the response and the predictors. For this so-called combination model we have all of the conditions present in models 2 to 5 together in one model. This case is crucial to showing the effectiveness of our methods since it is not always know in advance in which ways data may deviate from linearity. We see in Tables 5.11 and 5.12 that the lowest false negative rates come from the stepwise and penalized methods when all transformations are used.

As with the previous models, LASSO frequently returns a set of zero variables, showing yet again that the model free permutation test and cross validation procedure are essential to performing model-free variable selection using OLS-based methods.

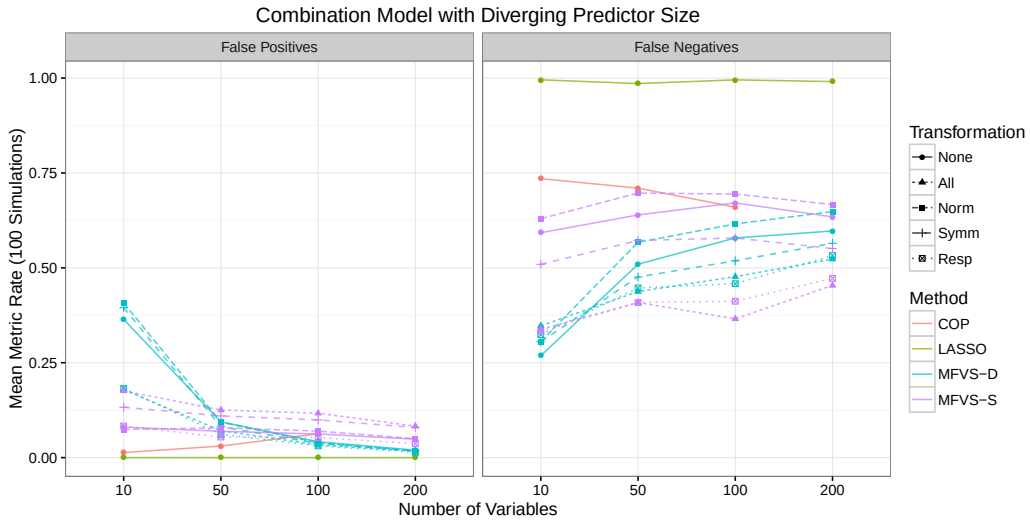

Figure 5.26: Combination model with diverging predictor size: mean false positive and false negative rates over 100 simulations.

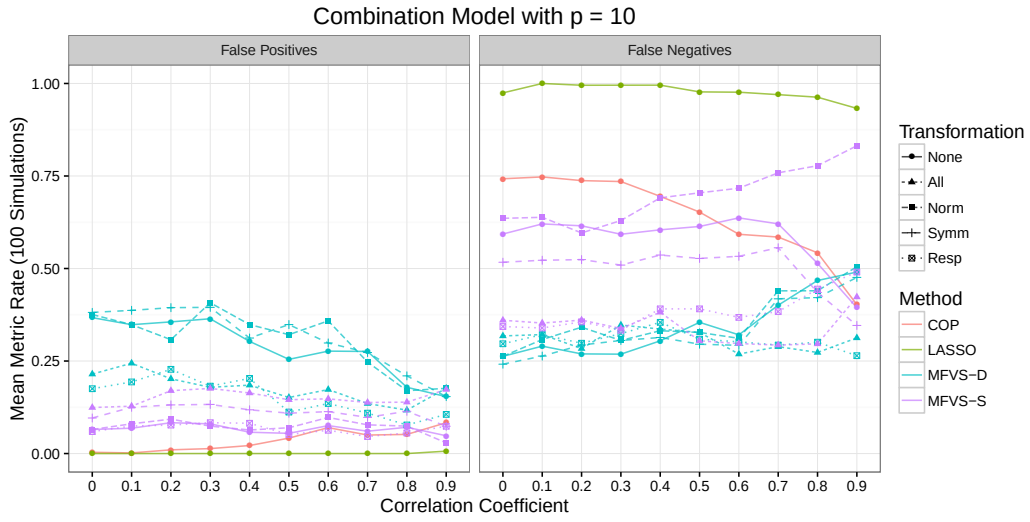

Figure 5.27: Combination model with varying correlation and $p = 10$: mean false positive and false negative rates over 100 simulations.

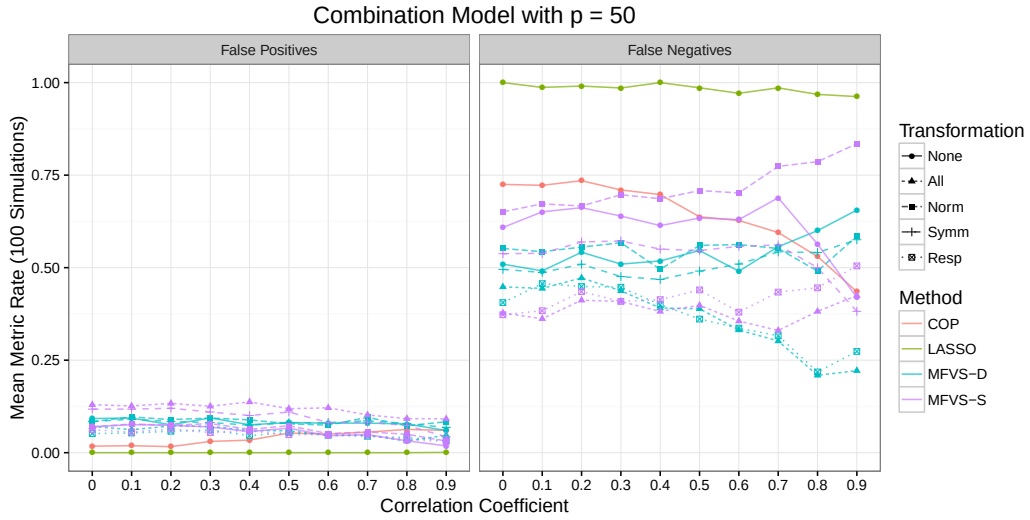

Figure 5.28: Combination model with varying correlation and $p = 50$: mean false positive and false negative rates over 100 simulations.

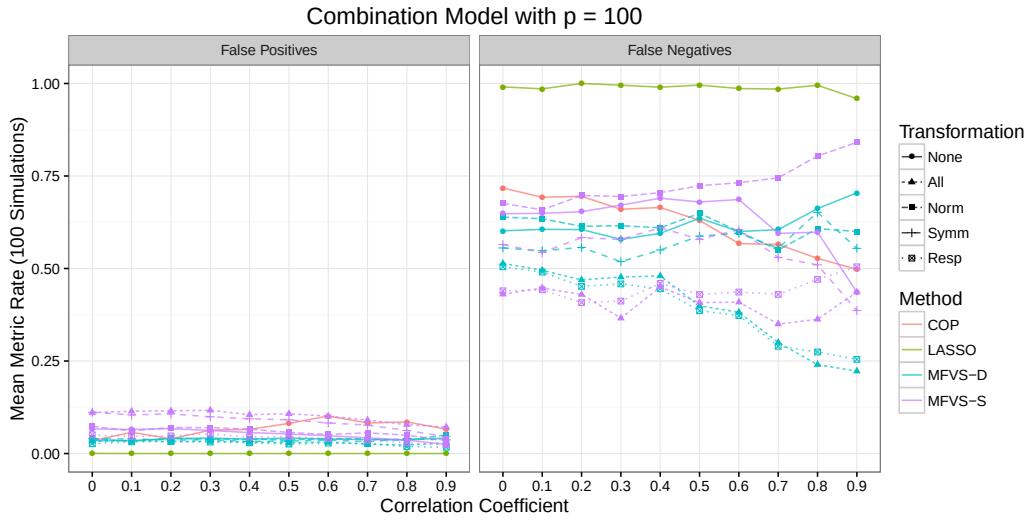

Figure 5.29: Combination model with varying correlation and $p = 100$: mean false positive and false negative rates over 100 simulations.

Method	f_1	f_2	f_3	f_4	FPR	FNR	$\lvert\mathcal{F}\rvert$
MFVS-D (All)	0.40	0.75	1.00	0.10	0.07	0.44	5.52
MFVS-D (None)	0.19	0.81	0.85	0.12	0.09	0.51	6.29
MFVS-D (Norm)	0.13	0.73	0.75	0.12	0.09	0.57	6.04
MFVS-D (Symm Link)	0.44	0.71	1.00	0.06	0.06	0.45	5.04
MFVS-D (Response)	0.15	0.90	0.94	0.10	0.08	0.48	5.85
MFVS-S (All)	0.46	0.73	1.00	0.17	0.13	0.41	8.13
MFVS-S (None)	0.08	0.87	0.44	0.06	0.07	0.64	4.63
MFVS-S (Norm)	0.06	0.75	0.35	0.06	0.08	0.70	4.83
MFVS-S (Symm Link)	0.50	0.79	1.00	0.08	0.05	0.41	4.87
MFVS-S (Response)	0.10	0.87	0.62	0.13	0.11	0.57	6.77
Lasso	0.00	0.06	0.00	0.00	0.00	0.99	0.06
COP	0.04	1.00	0.08	0.04	0.03	0.71	2.56

Table 5.11: Combination model: simulation results for $p = 50$ and $\rho = 0.3$ for each method (transformations in parentheses).

Method	f_1	f_2	f_3	f_4	FPR	FNR	$\lvert\mathcal{F}\rvert$
MFVS-D (All)	0.22	0.63	1.00	0.06	0.02	0.52	5.11
MFVS-D (None)	0.09	0.70	0.80	0.02	0.02	0.60	5.44
MFVS-D (Norm)	0.06	0.67	0.69	0.00	0.02	0.65	4.81
MFVS-D (Symm Link)	0.22	0.61	1.00	0.04	0.01	0.53	4.59
MFVS-D (Response)	0.06	0.72	0.91	0.06	0.02	0.56	5.19
MFVS-S (All)	0.33	0.74	1.00	0.11	0.08	0.45	18.43
MFVS-S (None)	0.11	0.78	0.56	0.02	0.05	0.63	10.96
MFVS-S (Norm)	0.09	0.74	0.44	0.06	0.05	0.67	11.09
MFVS-S (Symm Link)	0.35	0.76	1.00	0.00	0.04	0.47	9.31
MFVS-S (Response)	0.17	0.80	0.74	0.09	0.08	0.55	17.37
Lasso	0.00	0.04	0.00	0.00	0.00	0.99	0.04
COP	0.00	0.00	0.00	0.00	NA	NA	1.00

Table 5.12: Combination model: simulation results for $p = 200$ and $\rho = 0.3$ for each method (transformations in parentheses).

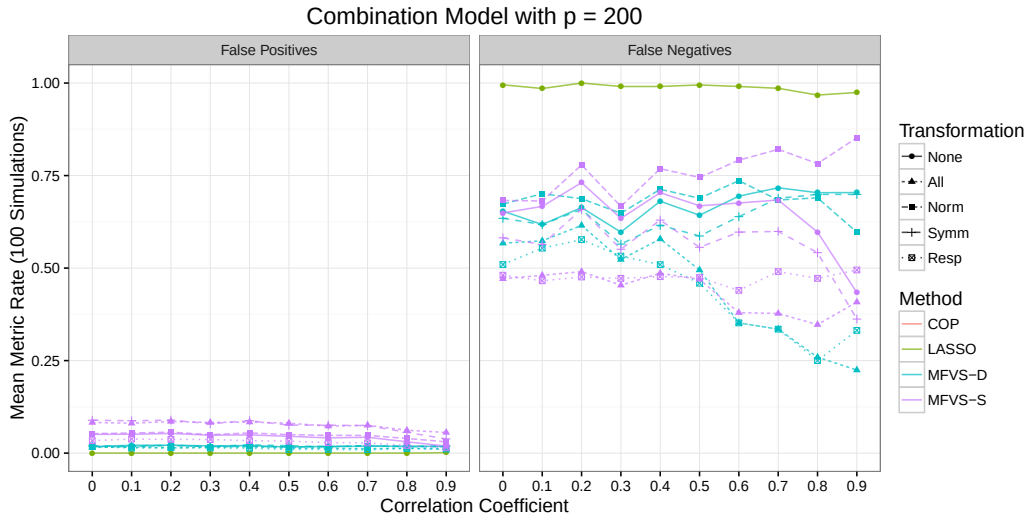

Figure 5.30: Combination model with varying correlation and $p = 200$: mean false positive and false negative rates over 100 simulations.

5.2 Analysis of Real Data

In this section we show the performance of our methods in two cases using real data. First we show the performance in the $n > p$ setting with an application from experimental psychology using MFVS-S. Second we show the performance in the $p > n$ setting with an application from genetics using MFVS-D.

5.2.1 The Good Judgment Project Data

To examine the performance of our methods on real world data, we first look to the field of experimental psychology. In 2015 Mellers et al. examined the forecasting accuracy of individuals in the domain of world politics. The researchers sought to understand forecasting accuracy by selecting the attributes of individual forecasters that were most associated with accurate forecasting.

The data came from the Good Judgment Project forecasting tournament where one goal was to understand how to elicit the most accurate forecasts from a large pool of forecasters. In the tournament accuracy was measured using a version of squared error loss called Brier score.

In 2015 Mellers, et al. found seven variables to be most highly correlated with forecasting accuracy and used those seven variables to build a model to predict forecasting accuracy. While performing variable selection by choosing the most correlated variables is an acceptable approach, there is some evidence

Selection Method	Training MSE	Prediction MSE	N Vars Selected
MFVS-S	0.0022	0.0032	7
COP	0.0017	0.0045	13
Mellers	0.0024	0.0037	7
LASSO	NA	NA	0

Table 5.13: Prediction Mean Squared Error on testing dataset and the number of variable selected for the Good Judgment Project dataset.

of nonlinearity and heterosckdasticity in the data, so a more refined approach may yield more accurate results. To further the previous work we will select from the same variables from the Good Judgment Project forecasting tournament using MFVS-S. We will examine the data from n=130 forecasters across 41 variables that may possibly relate to forecasting accuracy. The variable are comprised of a mix of demographic information, forecasting behavior and psychological test results. For more information on the variables used see Mellers et al. (2015).

For variable selection we will again use MFVS-S (with all three transformations), LASSO, and COP. Then we will build a generalized additive model using splines for each set of selected variables and we will build an additional model using the original seven variables identified by Mellers et al. To ensure that we are not overfitting we use 80% of the data for model building and 20% of the data for testing. The final judge of accuracy will be the prediction mean squared error on the testing set.

	COP	LASSO	Mellers	MFVS-S
Age	✓			
Open Minded Thinking Score			✓	
Total Forecasts Made	✓			
News Articles Read	✓			✓
Political Knowledge Score			✓	
China Political Knowledge Score	✓			
Russia Political Knowledge Score	✓			
CRT Score	✓		✓	✓
Finance Score	✓			
GPK Score	✓			
Helping Score				✓
Motivational Prosocial Score	✓			
Avg. Forecasts per Question	✓		✓	✓
Need for Cognition Score	✓			
Numeracy Score			✓	✓
Ravens Score	✓		✓	
Time Contribution				✓
Question Viewtime	✓		✓	✓
Number Selected	13	0	7	7

Table 5.14: Details of which variables were selected by which method for the Good Judgment Project dataset.

MFVS-S selects 7 variables, the same number as Mellers et al., but Table 5.2.1 shows that only 4 of the 7 variables selected by Mellers are also selected by MFVS-S. We believe the addition of these variable furthers the understanding of the dataset. COP selects a total of 17 variables. Given COP's tendency towards a high rate of false positives it is not surprising to see such a large number of selected variables and yet a higher prediction mean squared error. LASSO selects zero predictors. This is consistent with the simulation studies where LASSO frequency had very high rates of false negatives under departure from a linear model. Finally Table 5.2.1 shows that MFVS-S has the lowest prediction MSE which provides evidence that MFVS-S is the most accurate variable selection method in this case.

5.2.2 Diffuse large B-cell lymphoma

Next we choose the diffuse large B-cell lymphoma dataset (Rosenwald et al., 2002) to investigate the performance of MFVS-D in the ultra-high dimensional setting. We selected this dataset as it has been shown in existing model-free literature to have evidence of nonlinearities (Zhu et al., 2011). This dataset consists of 7399 gene expression levels ($p = 7399$) for 240 patients ($n = 240$). The response variable is patient survival time. Since it is computationally infeasible to perform any of the variable selection methods on 7,399 variables

with only 240 observations, we use distance correlation screening (Li, 2012) to reduce the variables to 300 before performing variable selection.

We select variables using MFVS-D with each of the possible transformations, no transformation, all of the transformations and some combinations of transformations. We also select variables using LASSO. Then we build a linear model for each set of selected variables and a generalized additive model (GAM) for each of the selected variables. The purpose of building two sets of models is to confirm whether or not the data has nonlinearities and thus the model free procedure is appropriate. We find in Table 5.2.2 that the GAM model with response has the lowest testing root MSE, indicating the presence of variables in the variance component or possibly another type of nonlinearity.

We also see that LASSO selects 43 variables which is a very high number; in fact LASSO returns too many variables to be able to use GAM. We are surprised by the large number of variables returned by LASSO. In the simulations and in the Good Judgment Project dataset analysis we saw that LASSO was more likely to return no variables at all, rather than return a large set of false positives. However we believe that this highlights the importance of the use of GAM in the cross validation step. We believe that the nonlinearities in the dataset are the cause for the very large set of variables returned by LASSO, since the LASSO cross validation was performed using the typical linear model error. Zhou et al. (2011) explain that only a small amount of genes will show

		GAM		Linear	
Selection Method	N	Train.	Pred.	Train.	Pred.
MFVS-D (None)	2	3.78	4.18	3.75	4.17
MFVS-D (All)	12	3.66	4.85	3.20	5.00
MFVS-D (Norm)	13	3.76	4.71	3.89	4.86
MFVS-D (Norm & Symm)	13	3.52	4.73	3.92	5.45
MFVS-D (Response)	7	3.75	4.05	4.00	4.12
MFVS-D (Norm & Resp)	5	3.89	4.65	3.86	4.69
LASSO	43	NA	NA	3.21	6.03

Table 5.15: Diffuse large B-cell lymphoma analysis results showing root mean squared error on training and testing sets using a linear model and a GAM model. For the MFVS-D method, the transformations used are in parentheses.

a true relationship with survival time and so we may presume that LASSO has in fact returned a very high number of false positives.

CHAPTER 6

A New Procedure for Model-free Interaction Detection

6.1 Review of variable selection methods with interaction detection

When dealing with a large number of predictor variables, the question of how to perform variable selection has been asked for a long time and many solutions have been put forth. A more complicated problem, with fewer potential solutions is the question of how to perform variable selection in the presence of interacting variables.

Before discussing any actual variable selection techniques for interaction detection, we must address the issue of hierarchical interactions. A model that includes a hierarchical interaction will have at least one main effect of the interaction pair also included in the model. Nelder (1977) postulated that a linear model with no main effects was "of no practical interest". His arguments were mathematical in nature and specific to linear models. Regarding polynomial regression models, Peixoto (1987) asserted that hierarchy must be preserved in a model since when considering two models where the variables have been linearly transformed, only the hypothesis tests concerning the highest order variables will be equivalent in each model, while the lower order variables will not. Peixoto's assertions extend to any higher order terms, not just interaction terms; i.e. if X_1^2 is included in the model, then X_1 must be as well. However, the author does concede that including higher-order terms but excluding their corresponding lower order terms may be acceptable when transformation is not desired and the model is used for prediction but not estimation. On the subject of variable selection, Peixoto recommends that variable selection algorithms restrict their search to, as he describes it, hierarchically well formulated models. Next we review some popular interaction detection methods.

Penalized Least Squares Methods. As an extension of the classical linear regression model, Choi, Li and Zhu (2010) considered the strong heredity

interaction model (SHIM)

$$Y = \beta_0 + \sum_j \beta_j X_j + \sum_{j>k} \delta_{jk} X_j X_k + \varepsilon, \tag{6.1}$$

and proposed a novel penalty to guarantee the strong hierarchical structure, i.e. both main effects in a two-way interaction are present in the model. Bien et. al (2013) proposed the hierarchical LASSO and considered both the strong hierarchical structure and the weak hierarchical structure, where only one main effect in a two-way interaction is present in the model.

Model (6.1) is a linear model of both the original predictors and their two-way interactions, and the methods introduced above will fail when this linear model assumption is violated. To relax the linear model assumption, Radchenko and James (2010) proposed "Variable selection using Adaptive Nonlinear Interaction Structures in High dimensions" (VANISH) under the following model

$$Y = \beta_0 + \sum_j f_j(X_j) + \sum_{j>k} f_{jk}(X_j X_k) + \varepsilon. \tag{6.2}$$

After expanding f_j and f_{jk} as linear combinations of basis functions, the nonlinear model (6.2) becomes similar to the linear model (6.1), and penalized least squares approaches can be implemented similar to Choi, Li and Zhu (2010) and Bien et al. (2013).

Exhaustive Search Methods. These methods exhaustively test every possible pair of interacting variables for significance in some measurement com-

parison to the response variable. Unlike the penalized least squares methods, exhaustive search methods do not necessarily estimate a model, but rather return a set of variables to be used for modeling. Exhaustive search is time-consuming and vulnerable to a large number of false positives. Many exhaustive search methods thus focus on reducing computation time (Goudey et al., 2013) as well as controlling the false positive rate (Wei et al., 2010). If well executed, exhaustive search methods can find potentially interacting variables even in the ultrahigh dimensional setting.

Sliced Inverse Regression with Interaction Detection. As we discussed before in Section 2.1, correlation pursuit (Zhong et al., 2012) can be viewed as a first-order method for variable selection with roots in sufficient dimension reduction. Instead of modeling $\mathrm{E}(X_j \mid Y)$ to test the significance of X_j as in correlation pursuit, Jiang and Li (2014) propose sliced inverse regression with interaction detection (SIRI) as a second-order method. SIRI achieves interaction detection through modeling $\mathrm{var}(X_j \mid Y)$. A likelihood ratio test is then proposed to determine whether X_j should be included in the model given that $\mathbf{x}_{\mathcal{F}}$ is already included. Like correlation pursuit, SIRI is applicable in the challenging ultrahigh dimensional setting when p is much larger than n. Unlike the penalized methods and exhaustive search methods described above, due to its inverse nature SIRI has no mechanism to differentiate different types of second order effects.

6.2 Revisit the pivotal quantity $\delta_{j|\mathcal{F}}$ in stepwise regression

For working index set $\mathcal{F}$ and candidate index $j \in \mathcal{F}^c$, the stepwise regression hypotheses from Section 2.1 consider

$$H_0 : Y \perp\!\!\!\perp X_j \mid \mathbf{x}_\mathcal{F} \text{ v.s. } H_a : Y \text{ is not independent of } X_j \text{ given } \mathbf{x}_\mathcal{F}.$$

The pivotal quantity is $\delta_{j|\mathcal{F}} = \mathrm{E}(YR_{j|\mathcal{F}})$, where $R_{j|\mathcal{F}} = X_j - \mathrm{E}(X_j \mid \mathbf{x}_\mathcal{F})$. We see from Proposition 2.1 that $\delta_{j|\mathcal{F}} = 0$ under H_0. The stepwise regression procedure will thus be effective if $\delta_{j|\mathcal{F}} \neq 0$ under H_a. The stepwise regression procedure will fail in the cases when $\delta_{j|\mathcal{F}} = 0$ under H_a. We use the following example to fix the idea.

Example 6. Let $\mathbf{x} \sim N(\mathbf{0}, \mathbf{I_p})$ and $Y = X_1X_2 + \varepsilon$. Suppose $\mathcal{F} = \emptyset$, $j = 1$. It follows that $\delta_{j|\mathcal{F}} = \mathrm{E}(YX_1) = 0$. In this case H_a is true because Y is not independent of X_1, and we have shown that $\delta_{j|\mathcal{F}} = 0$ under H_a. Using similar argument, we can see that $\delta_{j|\mathcal{F}} = 0$ for $\mathcal{F} = \emptyset$ and $j = 2$. The stepwise regression algorithm thus fails to add either X_1 or X_2 in the first addition step. Now consider a modified example. Let $\mathbf{x} \sim N(\mathbf{0}, \mathbf{I_p})$ and $Y = X_1 + X_1X_2 + \varepsilon$. In the initial step, consider $\mathcal{F} = \emptyset$. Then $\delta_{j|\mathcal{F}} = 0$ for $j = 2, \ldots, p$, and $\delta_{1|\mathcal{F}} = \mathrm{E}(YX_1) = 1$. Hence the working index set $\mathcal{F}$ will be updated to $\mathcal{F} = \{1\}$. In the next step, consider $\mathcal{F} = \{1\}$ and $j = 2, \ldots, p$. We have $\delta_{j|\{1\}} = \mathrm{E}(YX_j) = 0$ for $j = 2, \ldots, p$. The working index set $\mathcal{F}$ will not be

updated, and the stepwise regression algorithm will fail to add X_2 given that X_1 is already in the model.

Consider model (6.2) with the strong hierarchical structure, where all significant predictors due to the interactions have to appear in the main effects as well. The insight from Example 6 is that stepwise regression based on $\delta_{j|\mathcal{F}}$ still works well at recovering $\mathcal{A}_0$ due to the strong hierarchical structure. In the modified example $Y = X_1 + X_1X_2 + \varepsilon$ with weak hierarchical structure, stepwise regression based on $\delta_{j|\mathcal{F}}$ will fail to recover the active predictor that appears in the interaction and does not appear in the main effect. Due to this observation, we drop the terms with the main effects in model (6.2), and consider the following interaction-only model

$$Y = \sum_{j>k} f_{jk}(X_jX_k) + \varepsilon. \tag{6.3}$$

Our goal is to determine the active predictors in model (6.3) without estimating the unknown link functions $f_{jk}(\cdot)$. For the interaction-only model (6.3), we remark that the stepwise regression procedure based on $\delta_{j|\mathcal{F}} = 0$ will totally fail.

In applications, there is strong evidence (Ritche et. al., 2001; Nagel, 2005; Jiang and Neapolitan, 2012) that the field of genomics is rife with situations where interacting genes are not present in the form of main effects. An imagined example of this would be a disease that only occurs when two genes are both expressed, but not if only one or neither is expressed. The inves-

tigation into the quantitative connection between disease and genetics is the motivating force behind the development of many recent interaction detection methods. Specifically, the phenomena of interacting genes is called epistasis. One may refer to Cordell (2002) for a thorough review of the statistical challenges in identifying epistasis. Motivated from the evidence in genetics studies that the interacting genes could be absent in the form of main effects, and that the interaction effects can be nonlinear (Moore, 2006), we focus on the interaction-only model (6.3) throughout this section.

6.3 Derivation of a new pivotal quantity

Under model (6.3) with $Y = \sum_{j>k} f_{jk}(X_j X_k) + \varepsilon$, our goal is to design a new test for

$$H_0 : Y \perp\!\!\!\perp X_j \mid \mathbf{x}_{\mathcal{F}} \text{ v.s. } H_a : Y \text{ is not independent of } X_j \text{ given } \mathbf{x}_{\mathcal{F}}.$$

Recall from Proposition 2.1 that $\delta_{j|\mathcal{F}}$ is connected to the first-order derivative of $\Psi_{\mathcal{F},j}(\mathbf{x}) = \mathrm{E}(Y \mid \mathbf{x}_{\mathcal{F}}, X_j)$. Motivated by the proof of Proposition 1.3, we now derive a new pivotal quantity based on the second-order derivative of $\Psi_{\mathcal{F},j}(\mathbf{x})$. Without loss of generality, we also assume $\mathrm{E}(Y) = 0$ for the following discussions.

Denote $\mathbf{x}_{\mathcal{F}\cup j} = (\mathbf{x}_{\mathcal{F}}^T, X_j)^T$ and $\mathbf{\Sigma}_{\mathcal{F}\cup j} = \mathrm{var}(\mathbf{x}_{\mathcal{F}\cup j})$. The Hessian matrix of $\Psi_{\mathcal{F},j}(\mathbf{x})$ is denoted as $\mathbf{H}(\Psi_{\mathcal{F},j}) \in \mathbb{R}^{(|\mathcal{F}|+1)\times(|\mathcal{F}|+1)}$. The element in the kth

row and ℓth column of $\mathbf{H}(\Psi_{\mathcal{F},j})$ is $h_{k\ell} = \frac{\partial^2 \Psi_{\mathcal{F},j}(\mathbf{x})}{\partial X^k_{\mathcal{F}\cup j} \partial X^\ell_{\mathcal{F}\cup j}}$, where $X^k_{\mathcal{F}\cup j}$ and $X^\ell_{\mathcal{F}\cup j}$ and the kth and ℓth elements in $\mathbf{x}_{\mathcal{F}\cup j}$ respectively. Consider the partition of $\mathrm{E}\{\mathbf{H}(\Psi_{\mathcal{F},j})\}$ as

$$\mathrm{E}\{\mathbf{H}(\Psi_{\mathcal{F},j})\} = \begin{pmatrix} \mathrm{E}\left\{\frac{\partial^2 \Psi_{\mathcal{F},j}(\mathbf{x})}{\partial \mathbf{x}_\mathcal{F} \partial \mathbf{x}_\mathcal{F}}\right\} & \mathrm{E}\left\{\frac{\partial^2 \Psi_{\mathcal{F},j}(\mathbf{x})}{\partial \mathbf{x}_\mathcal{F} \partial X_j}\right\} \\ \mathrm{E}\left\{\frac{\partial^2 \Psi_{\mathcal{F},j}(\mathbf{x})}{\partial X_j \partial \mathbf{x}_\mathcal{F}}\right\} & \mathrm{E}\left\{\frac{\partial^2 \Psi_{\mathcal{F},j}(\mathbf{x})}{\partial X_j \partial X_j}\right\} \end{pmatrix}. \tag{6.4}$$

Apply second-order Stein's Lemma and we have

$$\mathrm{E}\{\mathbf{H}(\Psi_{\mathcal{F},j})\} = \mathbf{\Sigma}^{-1}_{\mathcal{F}\cup j} \mathrm{E}(\Psi_{\mathcal{F},j} \mathbf{x}_{\mathcal{F}\cup j} \mathbf{x}^T_{\mathcal{F}\cup j}) \mathbf{\Sigma}^{-1}_{\mathcal{F}\cup j} - \mathbf{\Sigma}^{-1}_{\mathcal{F}\cup j} \mathrm{E}(\Psi_{\mathcal{F},j}). \tag{6.5}$$

Define $(\mid \mathcal{F} \mid +1) \times (\mid \mathcal{F} \mid +1)$ dimensional matrix $\mathbf{A}$ and the partition of $\mathbf{\Sigma}_{\mathcal{F}\cup j}$ as

$$\mathbf{A} = \begin{pmatrix} \mathbf{I}_{|\mathcal{F}|} & \mathbf{0} \\ -\mathrm{E}^T(\mathbf{x}_\mathcal{F} X_j) \mathbf{\Sigma}^{-1}_\mathcal{F} & 1 \end{pmatrix} \text{ and } \mathbf{\Sigma}_{\mathcal{F}\cup j} = \begin{pmatrix} \mathbf{\Sigma}_\mathcal{F} & \mathrm{E}(\mathbf{x}_\mathcal{F} X_j) \\ \mathrm{E}^T(\mathbf{x}_\mathcal{F} X_j) & \sigma^2_j \end{pmatrix},$$

where $\sigma^2_j = \mathrm{var}(X_j)$. It follows that $\mathbf{A}\mathbf{x}_{\mathcal{F}\cup j} = \{\mathbf{x}^T_\mathcal{F}, R_{j|\mathcal{F}}\}^T$ with $R_{j|\mathcal{F}} = X_j - \mathrm{E}^T(\mathbf{x}_\mathcal{F} X_j) \mathbf{\Sigma}^{-1}_\mathcal{F} \mathbf{x}_\mathcal{F}$. Furthermore, let $\sigma^2_{j|\mathcal{F}} = \mathrm{var}(R_{j|\mathcal{F}}) = \sigma^2_j - \mathrm{E}^T(\mathbf{x}_\mathcal{F} X_j) \mathbf{\Sigma}^{-1}_\mathcal{F} \mathrm{E}(\mathbf{x}_\mathcal{F} X_j)$ and we have

$$\mathbf{A}\mathbf{\Sigma}_{\mathcal{F}\cup j} = \begin{pmatrix} \mathbf{\Sigma}_\mathcal{F} & \mathrm{E}(\mathbf{x}_\mathcal{F} X_j) \\ \mathbf{0} & \sigma^2_{j|\mathcal{F}} \end{pmatrix} \text{ and } (\mathbf{A}\mathbf{\Sigma}_{\mathcal{F}\cup j})^{-1} = \begin{pmatrix} \mathbf{\Sigma}^{-1}_\mathcal{F} & -\sigma^{-2}_{j|\mathcal{F}} \mathbf{\Sigma}^{-1}_\mathcal{F} \mathrm{E}(\mathbf{x}_\mathcal{F} X_j) \\ \mathbf{0} & \sigma^{-2}_{j|\mathcal{F}} \end{pmatrix} \tag{6.6}$$

(6.5), (6.6), and the fact that $\mathrm{E}(\Psi_{\mathcal{F},j}) = \mathrm{E}\{\mathrm{E}(Y \mid \mathbf{x}_\mathcal{F}, X_j)\} = \mathrm{E}(Y) = 0$ together implies

$$\mathrm{E}\{\mathbf{H}(\Psi_{\mathcal{F},j})\} = \begin{pmatrix} \mathbf{\Sigma}^{-1}_\mathcal{F} & \mathbf{c} \\ \mathbf{0} & \sigma^{-2}_{j|\mathcal{F}} \end{pmatrix} \begin{pmatrix} \mathrm{E}(Y \mathbf{x}_\mathcal{F} \mathbf{x}^T_\mathcal{F}) & \mathrm{E}(Y \mathbf{x}_\mathcal{F} R_{j|\mathcal{F}}) \\ \mathrm{E}^T(Y \mathbf{x}_\mathcal{F} R_{j|\mathcal{F}}) & \mathrm{E}(Y R^2_{j|\mathcal{F}}) \end{pmatrix} \begin{pmatrix} \mathbf{\Sigma}^{-1}_\mathcal{F} & \mathbf{0} \\ \mathbf{c} & \sigma^{-2}_{j|\mathcal{F}} \end{pmatrix} \tag{6.7}$$

where $\mathbf{c} = -\sigma_{j|\mathcal{F}}^{-2}\boldsymbol{\Sigma}_{\mathcal{F}}^{-1}\mathrm{E}(\mathbf{x}_{\mathcal{F}}X_j)$. Calculate the matrix multiplication in (6.7) and compare with the partition of $\mathrm{E}\{\mathbf{H}(\Psi_{\mathcal{F},j})\}$ in (6.4). We get

$$\mathrm{E}\left\{\frac{\partial^2\Psi_{\mathcal{F},j}(\mathbf{x})}{\partial\mathbf{x}_{\mathcal{F}}\partial X_j}\right\} = \sigma_{j|\mathcal{F}}^{-2}\boldsymbol{\Sigma}_{\mathcal{F}}^{-1}\mathrm{E}(Y\mathbf{x}_{\mathcal{F}}R_{j|\mathcal{F}}) + \sigma_{j|\mathcal{F}}^{-2}\mathbf{c}\mathrm{E}(YR_{j|\mathcal{F}}^2). \tag{6.8}$$

The two key quantities on the right hand side of (6.8) are $\mathrm{E}(Y\mathbf{x}_{\mathcal{F}}R_{j|\mathcal{F}})$ and $\mathrm{E}(YR_{j|\mathcal{F}}^2)$. Note that $R_{j|\mathcal{F}} = X_j - \mathrm{E}(X_j \mid \mathbf{x}_{\mathcal{F}})$ is the residual from the regression between X_j and $\mathbf{x}_{\mathcal{F}}$. When we test $H_0 : Y \perp\!\!\!\perp X_j \mid \mathbf{x}_{\mathcal{F}}$, the term $\mathrm{E}(Y\mathbf{x}_{\mathcal{F}}R_{j|\mathcal{F}})$ can be used to test the effect on Y due to the interaction between X_j and $X_{\mathcal{F}}$, while $\mathrm{E}(YR_{j|\mathcal{F}}^2)$ can be used to test the second-order effect of X_j given that $\mathbf{x}_{\mathcal{F}}$ is already in the model. For our purpose of interaction detection in model (6.3), consider pivotal quantity $\mathbf{w}_{\mathcal{F},j} = \mathrm{E}(Y\mathbf{x}_{\mathcal{F}}R_{j|\mathcal{F}})$.

Proposition 6.1. *Suppose* $\mathbf{x}$ *is normal with* $E(\mathbf{x}) = \mathbf{0}$. *Then* $\mathbf{w}_{\mathcal{F},j} = \mathbf{0}$ *under* $H_0 : Y \perp\!\!\!\perp X_j \mid \mathbf{x}_{\mathcal{F}}$.

Proof. Under H_0, $Y \perp\!\!\!\perp X_j \mid \mathbf{x}_{\mathcal{F}}$ implies $\mathrm{E}(Y \mid \mathbf{x}_{\mathcal{F}}) = \mathrm{E}(Y \mid \mathbf{x}_{\mathcal{F}}, X_j)$. Note that

$$\mathrm{E}\{Y\mathbf{x}_{\mathcal{F}}\mathrm{E}(X_j \mid \mathbf{x}_{\mathcal{F}})\} = \mathrm{E}\{Y\mathrm{E}(X_j\mathbf{x}_{\mathcal{F}} \mid \mathbf{x}_{\mathcal{F}})\} = \mathrm{E}\{\mathrm{E}(Y \mid \mathbf{x}_{\mathcal{F}})X_j\mathbf{x}_{\mathcal{F}}\}.$$

Plug in $\mathrm{E}(Y \mid \mathbf{x}_{\mathcal{F}}) = \mathrm{E}(Y \mid \mathbf{x}_{\mathcal{F}}, X_j)$ and we get

$$\mathrm{E}\{Y\mathbf{x}_{\mathcal{F}}\mathrm{E}(X_j \mid \mathbf{x}_{\mathcal{F}})\} = \mathrm{E}\{\mathrm{E}(Y \mid \mathbf{x}_{\mathcal{F}}, X_j)X_j\mathbf{x}_{\mathcal{F}}\} = \mathrm{E}(Y\mathbf{x}_{\mathcal{F}}X_j).$$

It follows that $\mathbf{w}_{\mathcal{F},j} = \mathrm{E}(Y\mathbf{x}_{\mathcal{F}}R_{j|\mathcal{F}}) = E\{Y\mathbf{x}_{\mathcal{F}}(X_j - \mathrm{E}(X_j \mid \mathbf{x}_{\mathcal{F}}))\} = 0$ under H_0. □

6.4 The backward interaction detection algorithm

Suppose we have i.i.d. observations $\{(\mathbf{x}^{(i)}, Y^{(i)}), i = 1, \ldots, n\}$ from model (6.3). We propose the backward interaction detection (BIND) algorithm to recover the active set $\mathcal{A}_0$ in this section. We have seen in the previous section that $\mathbf{w}_{\mathcal{F},j} = \mathrm{E}(Y\mathbf{x}_{\mathcal{F}}X_{j|\mathcal{F}})$ is a pivotal quantity for testing

$$H_0 : Y \perp\!\!\!\perp X_j \mid \mathbf{x}_{\mathcal{F}} \text{ v.s. } H_a : Y \text{ is not independent of } X_j \text{ given } \mathbf{x}_{\mathcal{F}}.$$

Denote $\mathbf{\Sigma}_{\mathbf{w}}^{\mathcal{F},j} = \mathrm{var}(Y\mathbf{x}_{\mathcal{F}}R_{j|\mathcal{F}})$. Let $\hat{\mathbf{w}}_{\mathcal{F},j}$ and $\hat{\mathbf{\Sigma}}_{\mathbf{w}}^{\mathcal{F},j}$ be the sample estimators of $\mathbf{w}_{\mathcal{F},j}$ and $\mathbf{\Sigma}_{\mathbf{w}}^{\mathcal{F},j}$ respectively. We consider the the following sample level test statistic

$$D^{\mathcal{F},j} = n\hat{\mathbf{w}}_{\mathcal{F},j}^T \left(\hat{\mathbf{\Sigma}}_{\mathbf{w}}^{\mathcal{F},j}\right)^{-1} \hat{\mathbf{w}}_{\mathcal{F},j}. \tag{6.9}$$

The asymptotic null distribution of $D^{\mathcal{F},j}$ is needed for testing the hypotheses.

From Proposition 4.1, we have seen that $R_{j|\mathcal{F}} \perp\!\!\!\perp (Y, \mathbf{x}_{\mathcal{F}})$ under H_0. Similar to the permutation test approach described in Section 4.1, we can approximate the distribution of $D^{\mathcal{F},j}$ under H_0 through random permutation. More specifically, given i.i.d. observations $\{(\mathbf{x}^{(i)}, Y^{(i)}), i = 1, \ldots, n\}$, we calculate the residuals $R_{j|\mathcal{F}}^{(i)}$ from the regression between $\mathbf{x}_j$ and $X_{\mathcal{F}}$. Combine these residuals with the original sample to get new dataset $\{(\mathbf{x}_{\mathcal{F}}^{(i)}, Y^{(i)}, R_{j|\mathcal{F}}^{(i)}), i = 1, \ldots, n\}$. We fix $\mathbf{x}_{\mathcal{F}}^{(i)}$ and $Y^{(i)}$, and randomly permute $R_{j|\mathcal{F}}^{(i)}$. Suppose the new sample

after the kth permutation is $\{(\mathbf{x}_{\mathcal{F}}^{(i)}, Y^{(i)}, \tilde{R}_{j|\mathcal{F}}^{(i),k}), i = 1, \ldots, n\}$. Following (6.9), the test statistic based on the kth permuted sample is

$$\tilde{D}_k^{\mathcal{F},j} = \left(\tilde{\mathbf{w}}_{\mathcal{F},j}^k\right)^T \left(\tilde{\mathbf{\Sigma}}_{\mathbf{w}_k}^{\mathcal{F},j}\right)^{-1} \tilde{\mathbf{w}}_{\mathcal{F},j}^k,$$

where $\tilde{\mathbf{w}}_{\mathcal{F},j}^k$ and $\tilde{\mathbf{\Sigma}}_{\mathbf{w}_k}^{\mathcal{F},j}$ are the sample estimators of $\mathbf{w}_{\mathcal{F},j}$ and $\mathbf{\Sigma}_{\mathbf{w}}^{\mathcal{F},j}$ based on the kth permuted sample. We conclude that $\tilde{D}_k^{\mathcal{F},j}$ has the same distribution as $D^{\mathcal{F},j}$ under H_0. Repeat the random permutation K times, and the p-value corresponding to test statistic $D^{\mathcal{F},j}$ can be approximated by

$$p_I^{\mathcal{F},j} = \frac{1}{K} \sum_{k=1}^{K} I(D^{\mathcal{F},j} > \tilde{D}_k^{\mathcal{F},j}), \text{ where } I(\cdot) \text{ is the indicator function.}$$

Note that the subscript I implies this is the p-value for interaction detection.

Now we describe the backwards interaction detection (BIND) algorithm. Note that if we start from $\mathcal{F} = \emptyset$, the second-order derivative $\mathrm{E}\{\partial^2 \Psi_{\mathcal{F},j}(\mathbf{x})/\partial \mathbf{x}_{\mathcal{F}} \partial X_j\}$ will reduce to the first-order derivative $\mathrm{E}\{\partial \Psi_{\mathcal{F},j}(\mathbf{x})/\partial X_j\}$, which is not suitable for interaction detection as we have seen earlier. This suggests we consider the following backward algorithm.

1. Set initial set $\mathcal{F} = \Omega$.

2. Find index j_d such that $j_d = \arg\max_{j \in \mathcal{F}} p_I^{\mathcal{F}\backslash j,j}$. If $p_I^{\mathcal{F}\backslash j_d,j_d} > p_d$, update $\mathcal{F}$ to be $\mathcal{F}\backslash j_d$. If $p_I^{\mathcal{F}\backslash j_d,j_d} \leq p_d$, no predictor is deleted.

3. Repeat Step 2 until no predictor can be deleted.

Instead of the purely backward algorithm, we can incorporate the addition step and have the modified BIND algorithm as follows

1. Set initial set $\mathcal{F} = \Omega$.

2. In the deletion step, find index j_d such that $j_d = \arg\max_{j \in \mathcal{F}} p_I^{\mathcal{F}\setminus j, j}$. If $p_I^{\mathcal{F}\setminus j_d, j_d} > p_d$, update $\mathcal{F}$ to be $\mathcal{F}\setminus j_d$. If $p_I^{\mathcal{F}\setminus j_d, j_d} \leq p_d$, no predictor is deleted.

3. In the addition step, find index j_a such that $j_a = \arg\min_{j \in \mathcal{F}^c} p_I^{\mathcal{F}, j}$. If $p_I^{\mathcal{F}, j_a} < p_a$, update $\mathcal{F}$ to be $\mathcal{F} \cup j_a$. If $p_I^{\mathcal{F}, j_a} \geq p_a$, no predictor is added.

4. Iterate between Step 2 and Step 3 until no predictor can be deleted or added.

6.5 Numerical studies

There are very few methods which can detect interactions in the model-free setting when first order effects are not also present in the model without evaluating all pairwise comparisons. We will compare our method to Hierarchical LASSO (hLasso) and SIRI. In hLasso, the authors assume that the interaction variables have a hierarchical structure, meaning that at least one main effect is present in the model. We will see through simulations that, in practice, the method will correctly select the correct variables even when neither of the

interacting variables are present in the model as a main effect, but the method will always show that at least one of the variables is in fact a main effect and will also often estimate spurious second order effects. This is a disadvantage, but the larger disadvantage of the hierarchical LASSO when compared with our method is that when the model assumptions are broken, the hierarchical LASSO no longer achieves selection consistency, erroneously returning large sets of variables which do not belong.

SIRI, on the other hand, does not assume a hierarchical structure, but also does not have a way to differentiate between different types of second-order effects. The results of this is that SIRI will not be able to differentiate between a quadratic terms, a heteroscedastic errors and interaction terms.

We used the following models where $\mathbf{x}$ is normally distributed with mean $\mathbf{0}$ variance $\mathbf{\Sigma} = \{\sigma_{ij}\}$, where $\sigma_{ij} = \rho^{|i-j|}$ where ρ increase from 0.1 to 0.9 in increments of 0.1. The sample size is $n = 500$ and $p = 10$:

1. Model 1 - Linear: $Y = X_1 X_2 + \varepsilon$, with $\varepsilon \sim \mathcal{N}(0, 2)$

2. Model 2 - Sine: $Y = \sin(2X_1 X_2) + \varepsilon$, with $\varepsilon \sim \mathcal{N}(0, 2)$

3. Model 3 - Exponential: $Y = \exp(X_1 X_2) + \varepsilon$, with $\varepsilon \sim \mathcal{N}(0, 2)$

4. Model 4 - Second Order with No Interactions:

   $Y = X_1^2 + X_{10}\varepsilon$, with $\varepsilon \sim \mathcal{N}(0, 1)$

	BIND	hLasso	SIRI
Model 1			
FP	1.32 (0.16)	0.00 (0.00)	0.11 (0.03)
FN	0.00 (0.00)	0.00 (0.00)	1.14 (0.10)
Model 2			
FP	0.80 (0.11)	1.43 (0.12)	0.19 (0.04)
FN	1.80 (0.04)	1.10 (0.09)	1.95 (0.03)
Model 3			
FP	2.24 (0.13)	.17 (0.05)	0.04 (0.02)
FN	2.00 (0.00)	0.88 (0.10)	0.00 (0.00)
Model 4			
FP	0.08 (0.04)	1.41 (0.13)	1.94 (0.02)
FN	2.00 (0.00)	2.00 (0.00)	2.00 (0.00)
Model 5			
FP	1.06 (0.18)	0.25 (0.06)	0.04 (0.02)
FN	1.24 (0.13)	0.28 (0.07)	1.69 (0.07)

Table 6.1: Simulation results for three methods under five simulated datasets. The table lists the mean and (standard error) of the number of false positives (FP) and false negatives (FN) for 100 simulations for each method and model. The table shows selection only for interacting variables, i.e. if a method selected a variable based on its first-order relationship with the response, the variable was not included in the number of false positives.

5. Model 5 - Rational with Additional Interaction:

$$Y = X_1/(0.5 + (1.5 + X_2)^2) + X_3X_4 + \varepsilon, \text{ with } \varepsilon \sim \mathcal{N}(0, 0.25)$$

The simulations were run 100 times and the results for $\rho = 0.2$ are in Table (6.1).

Model 1 is a linear model. BIND performs with complete accuracy, selecting X_1 and X_2 and no other variables each time. Lasso also performs very well, but selects an additional pair of interacting variables in more than half of the simulations. Due to its hierarchical assumptions, Lasso also incorrectly selected an average of 4.12 first order predictors and 1.20 quadratic predictors

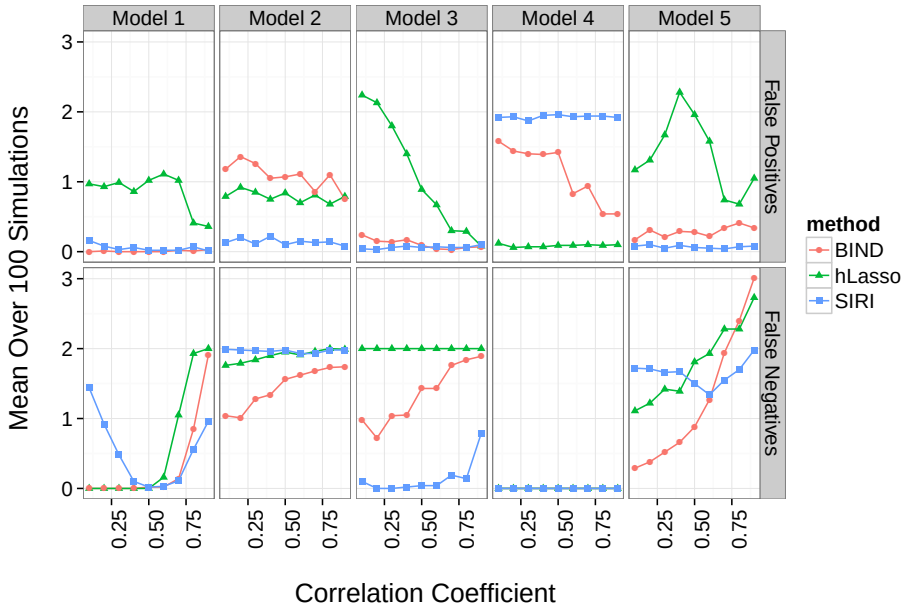

Figure 6.1: Mean (and standard error) of False-negatives (FN) and false-positives (FP) in 100 simulations for each interaction detection method and each simulated model.

on each simulated run of model 1. SIRI performs surprisingly poor on model 1, returning no selected variables in over half of the runs.

Model 2 is the most nonlinear of the five simulated models. While it does have the best performance of the three methods, BIND still only selects the correct predictors about half of the time. However, hLasso and SIRI miss the interacting variables almost all of the time. On average, hLasso returns 1.89 first order predators and 0.67 quadratic predictors for model 2. In model 3, another highly nonlinear model we see similar results to Model 3 with BIND and hLasso. However, SIRI performs nearly perfectly. Again, hLasso returns

additional variables with first and second-order effects, returning on average 2.47 and 1.98 of each type respectively.

Model 4 is included to illustrate the behavior of each method in the presence of a quadratic variable and heteroscedastic term, but no interacting variables. Since there are no interacting variables to detect, each method necessarily has no false negatives. BIND never returns the quadratic variable or the heteroscedastic variable, but does falsely return other variables erroneously and unpredictably. As expected, SIRI returns the quadratic variable in every run and the heteroscedastic variable in nearly all runs. Also as expected, hLasso identifies the quadratic term in nearly ever run but also returns a first-order effect with the quadratic term in order to enforce its hierarchical constraints; hLasso never returns the heteroscedastic term.

In model 5 BIND performs very well compared to the other methods, with a much lower false negative rate than either of the other two methods. Hierarchical Lasso tends to miss the terms in the rational part of the function and SIRI tends to miss half of the terms on average. Again, hLasso wrongly returns a high number of first and second-order effects at an average across all runs of 4.05 and 1.06 respectively.

These simulations show that interaction detection in a nonlinear setting is far from a trivial task. We see that BIND outperforms hLasso and SIRI in several situations, but still suffers from its own drawbacks, including poorer

performance under highly correlated predictors. Hierarchical Lasso tends to return larger numbers of variables and indicates linear and quadratic effects when none are present. SIRI, unfortunately, is unpredictable in its behavior and cannot differentiate between interacting variables and other second-order effects.

CHAPTER 7

Discussion

In this thesis we have shown that connection between sufficient dimension reduction and model-free variable selection can be developed in such a way as to extend traditionally linear methods to facilitate model-free variable selection. Today, in the world of ultra-high dimensional data and exploratory, rather than designed analysis, the importance of less restrictive variable selection techniques cannot be overstated.

In Section 2 we introduced model-free variable selection through OLS, developing a stepwise approach and a penalized approach to variable selection. In Section 3 we introduced three marginal transformations to extend the capabilities of the procedures introduced in Section 2. With the addition of a model free permutation test and cross validation procedure in Section 4, we

developed the sample-level algorithms that make model-free variable selection through OLS possible.

The simulations in Section 5 show empirically that it is possible to perform model free variable selection using the methods developed in Section 4. The simulations also show that directly applying a linear method to nonlinear data will lead to very poor results. We also see through the simulations that the addition of transformations serves to make our model-free procedures useful in a wider variety of cases.

Finally in Section 6 we introduced the concept of model-free interaction detection through PHD. We showed that our stepwise method can be used to select interacting terms in a model without any main effects. We also showed though simulations that while the task of selecting interacting variables in the model-free setting is quite challenge, our method performs very well.

REFERENCES

[1] Adragni, K. P. and Cook, R. D. (2009), "Sufficient Dimension Reduction and Prediction in Regression", Philosophical Transactions of the Royal Society A: Physical, Mathematical and Engineering Sciences, **367(1906)**, 4385–4405

[2] Bateson, W. (1909), "Mendel's Principles of Heredity", Cambridge University Press

[3] Bien, J., Taylor, J. and Tibshirani, R. (2013), "A lasso for hierarchical interactions", *The Annals of Statistics*, **41**(3), 1111–1141.

[4] Bondell, H. D., Li, L. (2009), "Shrinkage inverse regression estimation for model-free variable selection", *Journal of the Royal Statistical Society: Series B*, **71**(1), 287–299.

[5] Box, G. E. P. and Cox, D. R. (1964), "An analysis of transformations (with discussion)", *Journal of the Royal Statistical Society: Series B*, **26**, 211–252.

[6] Candes, E., Tao, T. (2007), "The Dantzig selector: Statistical estimation when p is much larger than n", *The Annals of Statistics*, **35 (6)**, 2313–2351.

[7] Chiang, A. P., et al. (2006). "Homozygosity mapping with SNP arrays identifies a novel gene for Bardet-Biedl syndrome (BBS10)". *Procedings of the National Academy of Sciences*, **103** 6287–6292.

[8] Choi, N. H., Li, W. and Zhu, J. (2010) "Variable selection with the strong heredity constraint and its oracle property", *Journal of the American Statistical Association*, **105**(489), 354–364.

[9] Cook, R. D. (1994), "Using dimension-reduction subspaces to identify important inputs in models of physical systems", In *Proc. Section on Physical and Engineering Sciences*, pp. 18–25. Alexandria, VA: American Statistical Association.

[10] Cook, R. D. (1998), "Regression Graphics: Ideas for Studying Regressions Through Graphics", New York: Wiley.

[11] Cook, R. D. (2007), "Fisher lecture: dimension reduction in regression (with discussion)", *Statistical Science*, **22**, 1–26.

[12] Cook, R. D., Li B. (2002), "Dimension reduction for conditional mean in regression", *The Annals of Statistics*, **30**(2) 465–474.

[13] Cook, R. D. (2004), "Testing predictor contributions in sufficient dimension reduction", *The Annals of Statistics*, **32**(3) 1062–1092.

[14] Cook, R. D. and Weisberg, S. (1991), "Discussion of sliced inverse regression for dimension reduction", *Journal of the American Statistical Association*, **86**, 328–332.

[15] Cordell, H. J. (2002), "Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans", *Human Molecular Genetics*, **11(20)**, 2463–2468.

[16] Cox, D. R. (1984), "Interaction", *International Statistical Review*, **52**, 1–31.

[17] Dong, Y., Yu, Z., and Zhu, L. P. (2014), "Model-Free Variable Selection for Conditional Mean in Regression" Technical report. Temple University.

[18] Fan, J. and Lv, J. (2008), "Sure independence screening for ultra-high dimensional feature space (with discussion)", *Journal of the Royal Statistical Society: Series B*, **70**, 849–911.

[19] Fan, J. and Li, R. (2001), "Variable Selection via Nonconcave Penalized Liklihood and its Oracle Properties", *Journal of the American Statistical Association*, **96**(456) 1348–1360.

[20] Fan, J. and Song, R. (2010), "Sure independence screening in generalized linear models with np-dimensionality", *The Annals of Statistics*, **38**(6) 3567–3604

[21] Goudey, B., et al. (2013), "GWIS - model-free, fast and exhaustive search for epistatic interactions in case-control GWAS", *BMC Genomics*, **14**(Suppl 3):S10

[22] Hesterberg, T., Choi, N. H., Meier, L., Fraley, C. (2008), "Least angle and L1 penalized regression: A review", *Statistics Surveys*, **2**, 61–93.

[23] Huang, J., Ma, S. and Zhang, C. H. (2008), "Adaptive Lasso for sparse high-dimensional regression models", *Statistica Sinica*, **18**(4), 1603–1618.

[24] James, J.M., Radchenko, P., Lv, J. (2009), "DASSO: Connections between the Datzig selector and lasso", *Journal of the Royal Statistical Society: Series B*, **71**(1), 1127–142.

[25] Jiang, B., Liu, J. S. (2014), "Variable selection for general index models via sliced inverse regression", *The Annals of Statistics*, **42**(5), 1751–1786.

[26] Jiang X., Neapolitan R. E. (2012), "Mining pure, strict epistatic interactions from high-dimensional datasets: ameliorating the curse of dimensionality". *PLoS ONE* **7**(10)

[27] Li, K-C. (1991), "Sliced Inverse Regression for Dimension reduction", *Journal of the American Statistical Association* **86**(414), 316–327

[28] Li, K-C. (1992), "On principal hessian directions for data visualization and dimension reduction: another application of Stein's lemma", *Journal of the American Statistical Association* **87**(420), 1025–1039.

[29] Li, L. (2007), "Sparse sufficient dimension reduction", *Biometrika* **93**(3), 603–613.

[30] Li, L., Cook, R. D., Nachtsheim, J. (2005), "Model-free variable selection" *Journal of the Royal Statistical Society: Series B*, **67**(2), 285–299.

[31] Li, K-C. and Duan, N. (1989), "Regression analysis under link violation", *The Annals of Statistics*, **17**, 1009–1052.

[32] Li, R., Zhong, W. and Zhu, L. (2012), "Feature screening via distance correlation learning", *Journal of the American Statistical Association* **107**, 1129–1139.

[33] Li, X., et al. (2014), "Flare: Family of Lasso Regression", *R package* version 1.5.0.

[34] Ma, Y. and Zhu, L. (2013), "A Review on Dimension Reduction", *International Statistics Review*, **81**, 134–150.

[35] Mai, Q. and Zou, H. (2015), "Nonparametric Variable Transformation in Sufficient Dimension Reduction", *Technometrics*, **57**(1), 1–10.

[36] Mellers, B., et al. (2015), "The psychology of intelligence analysis: Drivers of prediction accuracy in world politics", *Journal of Experimental Psychology: Applied*, **21**(1), 1–14.

[37] Moore, J. H., et al. (2006), "A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility", *Journal Theoretical Biology*, **241**, 252–261.

[38] Nagel, R.L. (2005), "Epistasis and the genetics of human diseases", *C. R. Biologies*, **328**(7), 606–615.

[39] Radchenko, P., James, G. M. (2010), "Variable selection using adaptive nonlinear interaction structures in high dimensions", *Journal of the American Statistical Association* **105**(492), 1541–1553.

[40] Ritchie, M. D. et al. (2001), "Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer", *The American Society of Human Genetics*, **69**(1), 138–147.

[41] Rosenwald, A. et al. (2002), "The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma", *The New England Journal of Medicine***346**(25): 1937–1947.

[42] Scheetz, T. E., Kim, A. K., Swiderski, R. E., Philp, A. R., Braun, T. A., Knudtson, K. L., Dorrance, A. M., DiBona, G. F., Huang,

J., Casavant, T. L., Sheffield, V. C., and Stone, E. M. (2006), "Regulation of gene expression in the mammalian eye and its relevance to eye disease" *Proceedings of the National Academy of Sciences* **103**(39) 14429–14434.

[43] Shin, Y.J. and Park, C.H. (2011), "Analysis of Correlation Based Dimension Reduction Methods," *International Journal of Applied Mathematics and Computer Science*, **21**(3), 549–558.

[44] Stein, C. (1981), "Estimation of the mean of a multivariate normal distribution," *The Annals of Statistics*, **9**, 1135–1151.

[45] Szekely, G.J., Rizzo, M.L., and Bakirov, N.K. (2007), "Measuring and Testing Dependence by Correlation of Distances", *The Annals of Statistics*, **35**(6), 2769–2794.

[46] Thompson, M. (1981), "Selection of variables in multiple regression: part I. A review and evaluation," *International Statistical Review*, **46**(1), 1–19.

[47] Tibshirani, R. (1996), "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B*, **58**, 267–288.

[48] Tsanas, A., Little, M. A., McSharry, M. E., Ramig, L.O. (2009), "Accurate telemonitoring of Parkinson's disease progression by non-invasive speech tests", *IEEE Transactions on Biomedical Engineering*, in press.

[49] Wang, T., et al. (2014), "Transformed sufficient dimension reduction", *Biometrika* **101**(4), 815–829.

[50] Wei, W. H., Knott, S., Haley, C. S., de Koning, D. J. (2010), "Controlling false positives in the mapping of epistatic QTL" *Heredity* **104**, 401–409.

[51] Yeo, IK., Johnson, R. (2000), "A New Family of Power Transformations to Improve Normality or Symmetry" *Biometrika***87**(4) 954–959.

[52] Yin, X. and Cook, R.D. (2002), "Dimension reduction for the conditional kth moment in regression" *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **64**(2)159–175.

[53] Yin, X. and Cook, R.D. (2003), "Estimating central subspaces via inverse third moments", *Biometrika*, **90**(1), 113–125.

[54] Yu, Z., Dong, Y., Zhu, L. (2015), "Trace pursuit: A general framework for model-free variable selection", *Journal of the American Statistical Association* (In Press)

[55] Zhang, C. (2010), "Nearly unbiased variable selection under minimax concave penalty", *The Annals of Statistics*, **38**(2), 894–942.

[56] Zhong, W., Zeng, P., Ma, P., Liu, J. S., Zhu, Y. (2005), "RSIR: regularized sliced inverse regression for motif discovery", *Bioinformatics*, **21**(22), 4169–4175.

[57] Zhong, W., Zhang, T., Zhu, Y. and Liu, J. S. (2012), "Correlation pursuit: forward stepwise variable selection for index models", *Journal of American Statistical Association*, **74**(5), 849–870.

[58] Zhu, L. P., Li, L., Li, R. and Zhu, L.X. (2011), "Model-free feature screening for ultrahigh dimensional data," *Journal of American Statistical Association*, **106**, 1464–1475.

[59] Zhu, L. P. and Zhu, L.X. (2009), "Dimension-reduction for conditional variance in regressions", *Statistica Sinica*, **19**, 869–883.

[60] Zou, H. (2006), "The adaptive Lasso and its oracle properties", *Journal of American Statistical Association*, **101**(476), 1418–1429.

[61] Zou, H. and Hastie, T. (2005), "Regularization and variable selection via the elastic net", *Journal of the Royal Statistical Society: Series B*, **67**(2), 301–320.